Multi-Modal Translation

Speech and Vision

Ondřej Bojar

May 7, 2020

EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Outline

- Overview of Multi-Modal Translation.
- Speech Translation $\approx$ ASR + MT.
    - Problems at ASR-MT boundary.
    - End-to-end SLT approaches.
- Visual information for MT.

Some pictures and tables from Sulubacak et al. (2019).

Overview of Multi-Modal MT

From survey by Sulubacak et al. (2019):

Overview of Multi-Modal MT

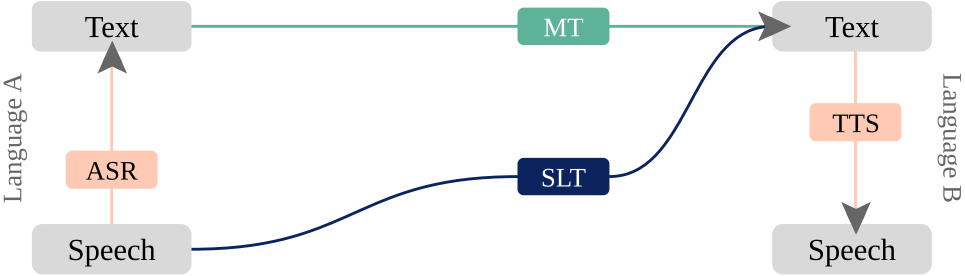
From survey by Sulubacak et al. (2019):

Overview of Multi-Modal MT

From survey by Sulubacak et al. (2019):

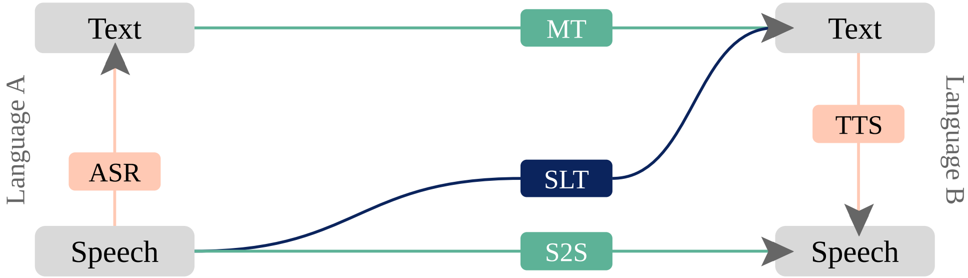
Overview of Multi-Modal MT

From survey by Sulubacak et al. (2019):



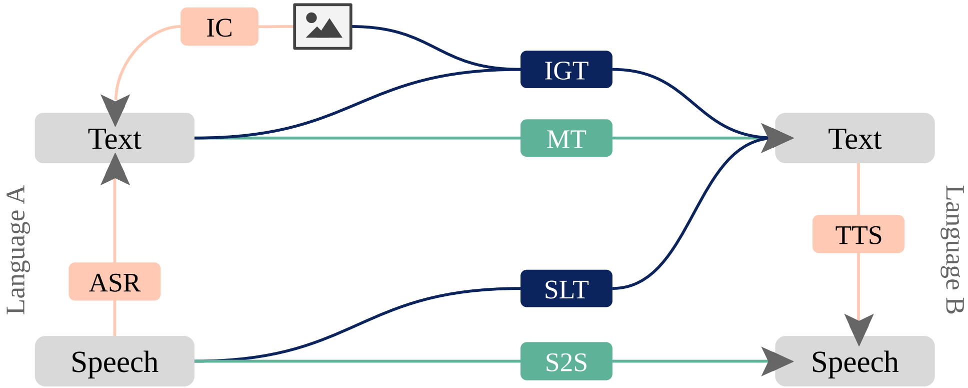
Overview of Multi-Modal MT

From survey by Sulubacak et al. (2019):



Overview of Multi-Modal MT

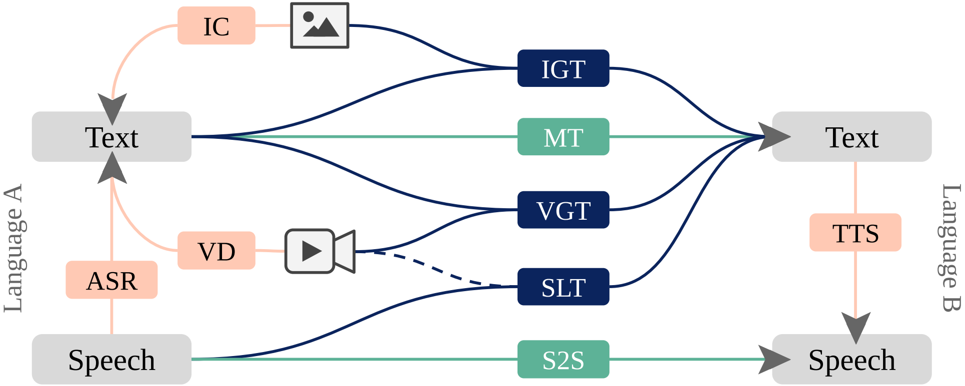
From survey by Sulubacak et al. (2019):



IGT = image-guided

Overview of Multi-Modal MT

From survey by Sulubacak et al. (2019):



IGT = image-guided, VGT = video-guided translation

Spoken Language Translation

Basic Terms

- MT = Machine Translation = Text Translation
  - Input are (mostly grammatically correct) individual sentences.
  - Sentences may come in documents or not.
  - (Document-level MT processes a sequence of sentences at once.)

Basic Terms

- MT = Machine Translation = Text Translation
  - Input are (mostly grammatically correct) individual sentences.
  - Sentences may come in documents or not.
  - (Document-level MT processes a sequence of sentences at once.)
- Incremental MT
  - MT of gradually growing input.
  - MT decides whether to wait for more words or emit current word.
  - Aims at stable output.

Basic Terms

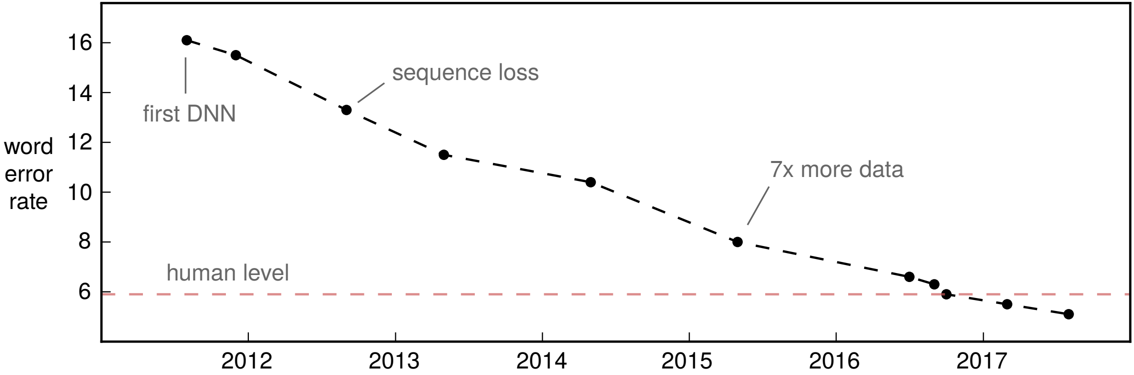
- MT = Machine Translation = Text Translation
  - Input are (mostly grammatically correct) individual sentences.
  - Sentences may come in documents or not.
  - (Document-level MT processes a sequence of sentences at once.)
- Incremental MT
  - MT of gradually growing input.
  - MT decides whether to wait for more words or emit current word.
  - Aims at stable output.
- SLT = Spoken Language Translation
  - Input is the sound in one language.
  - Output is text (sometimes also speech).
  - Sentences may or may not be assumed and produced.
- S2S = S2ST = Speech-to-Speech Translation
  - Direct modelling, e.g. can aim to preserve voice or prosodics.

Spoken Language Translation

Cascaded ASR + MT

NN Prospects: ASR Surpassing Humans

- Switchboard conversational speech benchmark (2000).
- 40 phone calls between two random native English speakers.



Plot by `https://awni.github.io/speech-recognition/`

MT Surpassing Humans for News

Seg-Level English→Czech 2018

	Ave. %	Ave. z	System
1	84.4	0.667	**CUNI-Transformer**
2	79.8	0.521	UEDIN
	78.6	0.483	Professional Translation
4	68.1	0.128	ONLINE-B
5	59.4	−0.178	ONLINE-A
6	54.1	−0.354	ONLINE-G

Doc-Aware English→German 2019

Ave.	Ave. z	System
90.3	0.347	**Facebook-FAIR**
93.0	0.311	Microsoft-WMT19-sent-doc
92.6	0.296	Microsoft-WMT19-doc-level
90.3	0.240	Professional Translation
87.6	0.214	MSRA-MADL
88.7	0.213	UCAM
89.6	0.208	NEU
87.5	0.189	MLLP-UPV
87.5	0.130	eTranslation
86.8	0.119	dfki-nmt
84.2	0.094	online-B
	... 10 more systems here ...	
76.3	−0.400	online-X
43.3	−1.769	en-de-task

See lecture #1 for all caveats of MT evaluation.

SLT Pipeline

1. Run ASR.
2. Run MT.

SLT Pipeline

1. ~~Run ASR~~ Recognize **lowercase words**.
2. ~~Run MT~~ Translate **sentences**.

SLT Pipeline

1. ~~Run ASR~~ Recognize **lowercase words**.
2. Segment into sentences.
3. ~~Run MT~~ Translate **sentences**.

SLT Pipeline

1. ~~Run ASR~~ Recognize **lowercase words**.
2. Segment into sentences.
3. Consider how to handle uncertainty!
4. ~~Run MT~~ Translate **sentences**.

SLT Pipeline

1. Acquire sound.
2. ~~Run ASR~~ Recognize **lowercase words**.
3. Segment into sentences.
4. Consider how to handle uncertainty!
5. ~~Run MT~~ Translate **sentences**.
6. Present output.

SLT Pipeline When Deployed

1. Acquire sound.
2. Ship to ASR worker.
3. ~~Run ASR~~ Recognize **lowercase words**.
4. Ship to sentence segmenter.
5. Segment into sentences.
6. Ship to translation worker.
7. ~~Consider how to handle uncertainty!~~
8. ~~Run MT~~ Translate **sentences**.
9. Ship to presentation worker.
10. Present output.

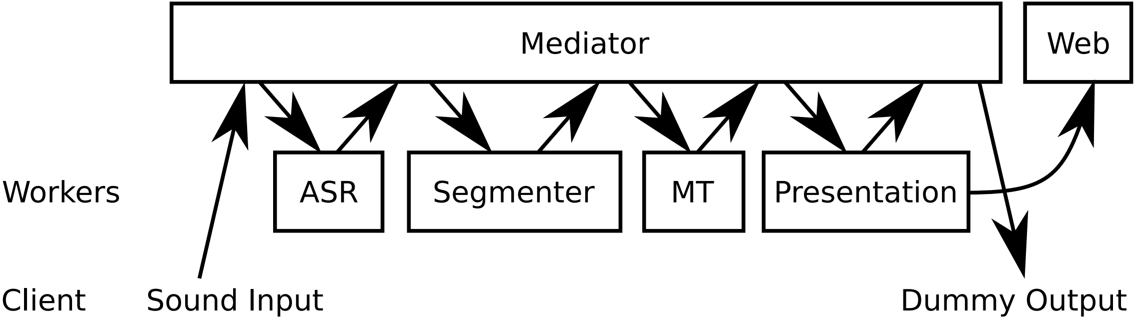
SLT Pipeline When Deployed

1. Acquire sound.
2. Ship to ASR worker.
3. ~~Run ASR~~ Recognize **lowercase words**.
4. Ship to sentence segmenter.
5. Segment into sentences.
6. Ship to translation worker.
7. ~~Consider how to handle uncertainty!~~
8. ~~Run MT~~ Translate **sentences**.
9. Ship to presentation worker.
10. Present output.

In realtime!

Overall Architecture in ELITR

- Components can run distributed, connected via “bi-sockets”.



- Connections always open, reused across clients.
- TCP communication $\Rightarrow$ relies on network capacity.

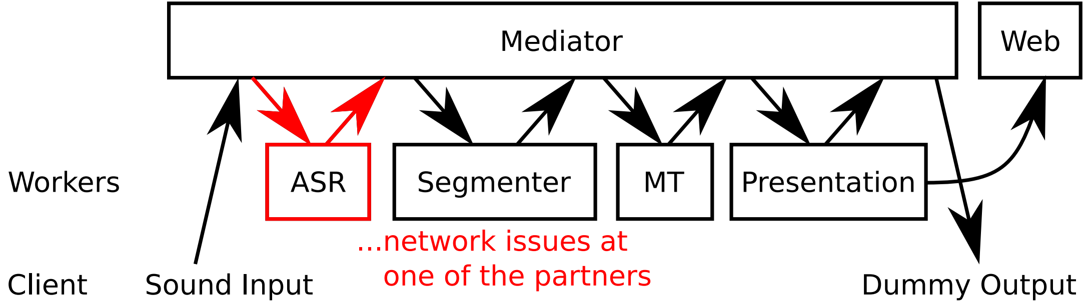
Spoken Language Translation

Network Issues

Failures Due to Setup

Over our test sessions, we saw:

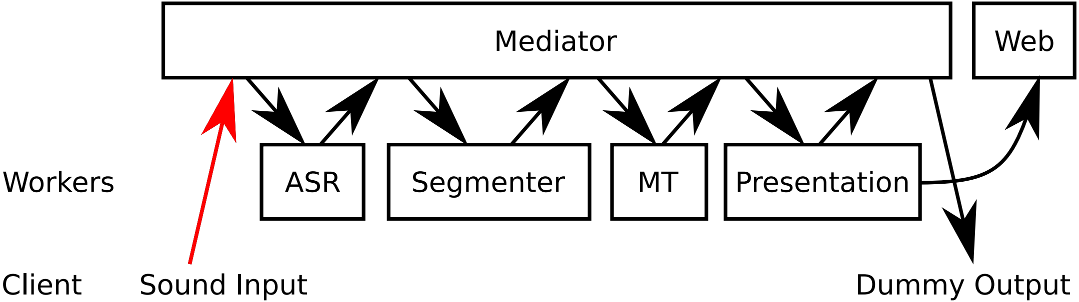
- slow network at various steps,
- partially working misconfiguration.



Failures Due to Setup

Over our test sessions, we saw:

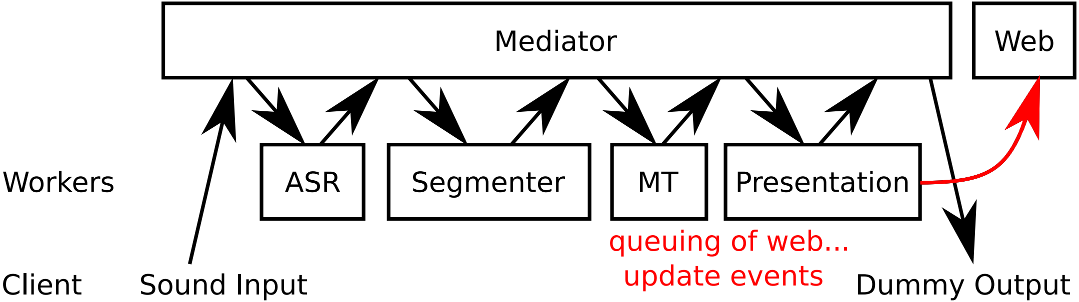
- slow network at various steps,
- partially working misconfiguration.



Failures Due to Setup

Over our test sessions, we saw:

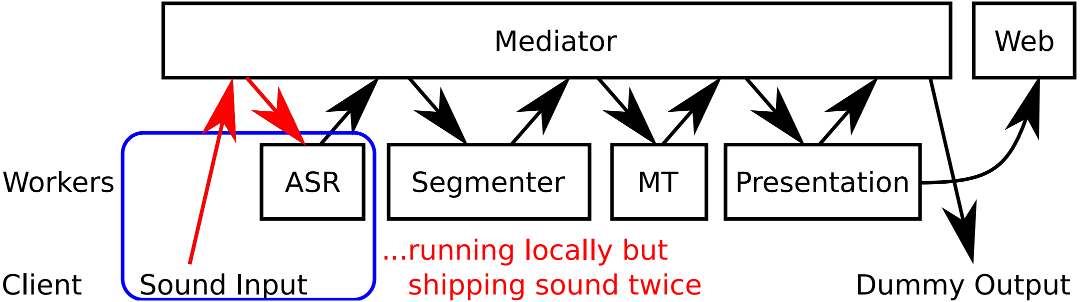
- slow network at various steps,
- partially working misconfiguration.



Failures Due to Setup

Over our test sessions, we saw:

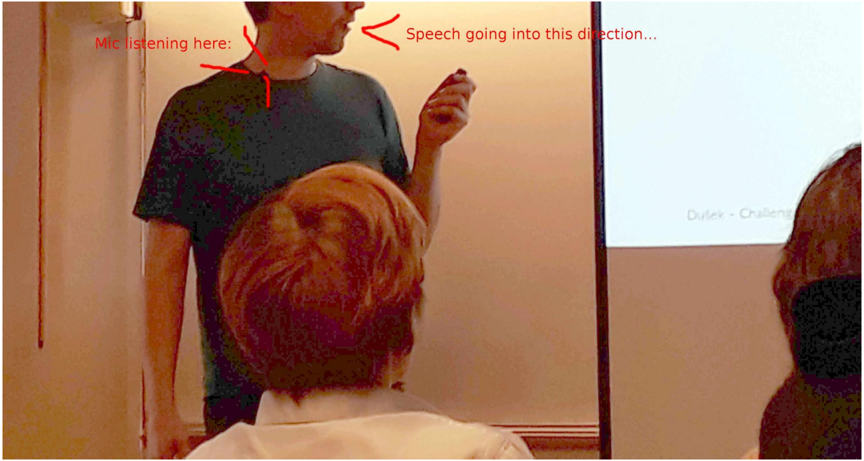
- slow network at various steps,
- partially working misconfiguration.



Spoken Language Translation

Sound Acquisition

Microphone Position



Headset Mic vs. Shirt Mic

A micro-test (just 3.5 minutes in total) with two microphones:

Word Error Rate	Headset	Shirt	Diff
EN ASR	**0.32**	0.39	-0.07
CS ASR	**0.14**	0.17	-0.03

Microphone Distance and Other Errors

https://www.sweetwater.com/insync/5-ways-your-mic-technique-is-ruining-your-vocals/

Microphone Distance and Other Errors

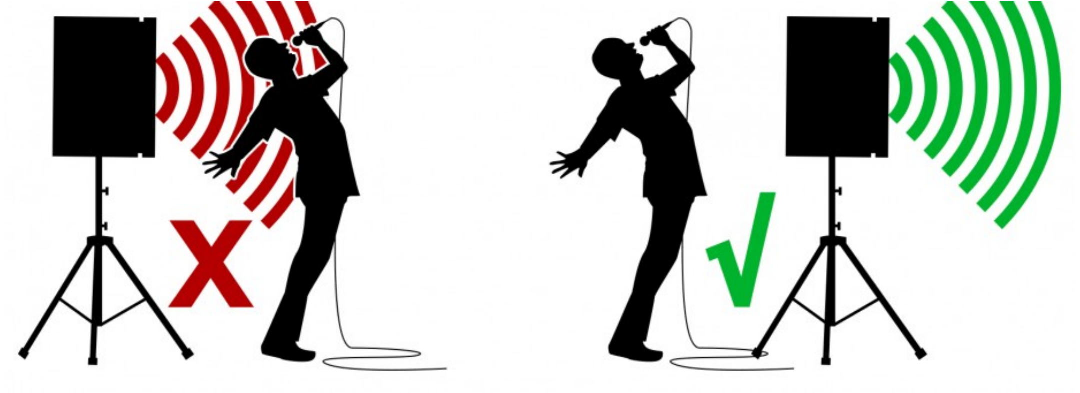
https://www.sweetwater.com/insync/5-ways-your-mic-technique-is-ruining-your-vocals/

Microphone Distance and Other Errors

https://www.sweetwater.com/insync/5-ways-your-mic-technique-is-ruining-your-vocals/

Microphone Distance and Other Errors

https://www.sweetwater.com/insync/5-ways-your-mic-technique-is-ruining-your-vocals/

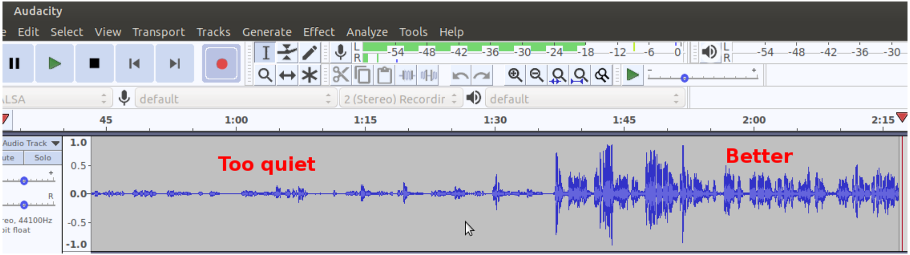


Volume Settings along the Pipeline

A number of volume controls is on the way:

- Wireless microphone output volume.
- Sound card input volume.
  - Line/Mic Level.
  - Padding.
- Automatic clipping of too loud signal.

⇒ You need to carefully 'track' the signal step by step.



Spoken Language Translation

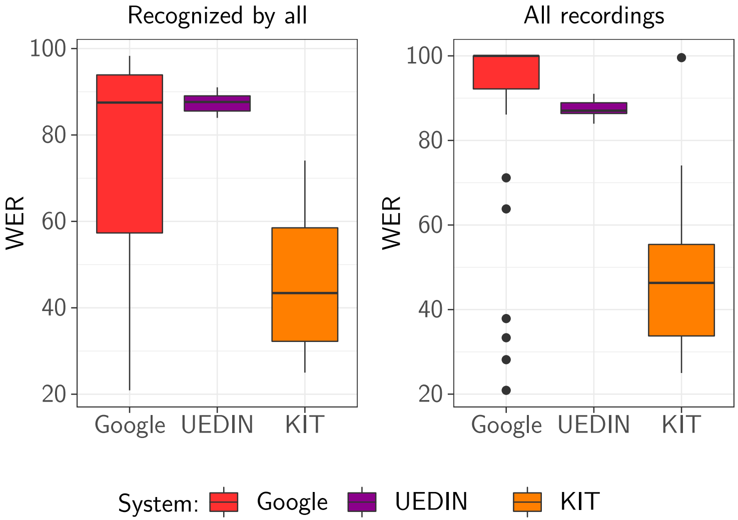
Realistic ASR Quality

ASR Challenges

Speaker intents	You have a **botel**? Oh, yes. We're situated in hearth of **České Budějovice**.
Reality	You have a bottle? Oh, yes. VeeR situated in haRd of **České Budějovice**. + BACKGROUND NOISE
Unknowledgeable person hears	You have a bottle? Oh, yes. We're situated in hearth of Che... WHICH CITY?
Noise-sensitive ASR	∅ oh yes the the of ∅
Noise-resistent ASR	you have somebody to Oh, yes, we are situated in hard which is can we do?
Knowledgeable person / Future ASR	You have a **botel**? Oh, yes, we're situated in hearth of **České Budějovice**.

- Non-"standard" pronunciation, background noise, OOV, named entities

ASR on Non-Native High-School Students



ASR of Non-Natives in Noisy Environment

- Human error level: 4–6% WER (word error rate).
- Best neural nets are reportedly there, too.
- Our test of 90-second speeches of high-school students:
  - Average WER: 40–50% KIT, 80–90% Google, UEDIN.

The best recognized segment:

Manual	Google	UEDIN	KIT
why do you wear those high heels , if you would wear some sneakers ? I know one really good store , that deals with the sale of freetime	why do ∅ where does high heels if you would wear some sneakers I no one really good star that deals with the sale of freedown food to	're ready where tells us if you would ∅ sneaker us I know won the really good store that deals with the sale of freedown food	why do are those highs heels if you would where some sneakers i no one really good story that deals with the sale of freetime food to our

Spoken Language Translation

Realistic MT Quality

General Translation Errors, Domain Issues

ASR	But it is much more difficult to ask if you do not have any clue.
MTde	Aber es ist viel schwieringer zu fragen, ob Sie keine Vorstellung davon haben.
MTcs	Je však mnohem těžší ptát se, zda nemáte ponětí.

- "if" should be translated as "wann"/"když" in this context.

ASR	You can be reported after some profanities.
MTcs	Můžete být hlášeni o některých profesních věcech.
Gloss	You can be reported due to some professional things.

ASR Errors Multiplied in MT

- Errors in ASR are mostly similar words.
  - Reasonably easy for the user to recover from transcript errors.
- MT takes these wrong words as fully trustworthy.
  - MT happily reorders the sentence to sound best, including wrong words.
  - No information about ASR and MT confidence available!

ASR	And the goal of my thesis is to fold.
MTcs	A cílem mé teorie je rozdrobit se.
Gloss	And the goal of my theory is to fall apart.
Ref	A cíle má moje teze dva.
Gloss	And there are two goals of my thesis.

Spoken Language Translation

ASR + MT Integration

ASR + MT Integration

- ASR emits string of lowercase words.
- MT expects individual correct sentences.

Options to bridge the gap:

1. Insert punctuation into ASR output $\Rightarrow$ new step: Segmentation.
2. Change ASR to predict directly correct punctuation.
3. Fully end-to-end SLT.

Approaches to Segmentation

- Language-Model-based: LM score without and with punctuation:

$$\mathrm{P}(\text{some sneakers I know}) \gtrless \mathrm{P}(\text{some sneakers, I know}) \gtrless$$
$$\gtrless \mathrm{P}(\text{some sneakers. I know}) \gtrless \mathrm{P}(\text{some sneakers? I know})$$

- Sequence-labelling:
  - Label each word with punctuation that should follow it.
  - Many techniques possible: HMM, CRF, LSTM, ...
- Machine-translation:
  - Input: Text without punctuation.
  - Output: Text with punctuation.
  - Approaches: PBMT, NMT.

A critical decision whether to allow access to the sound:

- Delays, prosody, intonation are very informative.

Errors in Segmentation

- Errors in precision lead to confusing MT output:

Speaker	...all too well...
ASR+Segm	...this approach does not generalize all too. Well, so to somehow concludes that the whole talk.

- Errors in recall make too much content unstable, see below.

Spoken Language Translation

End-to-End SLT

Motivation for End-to-End SLT

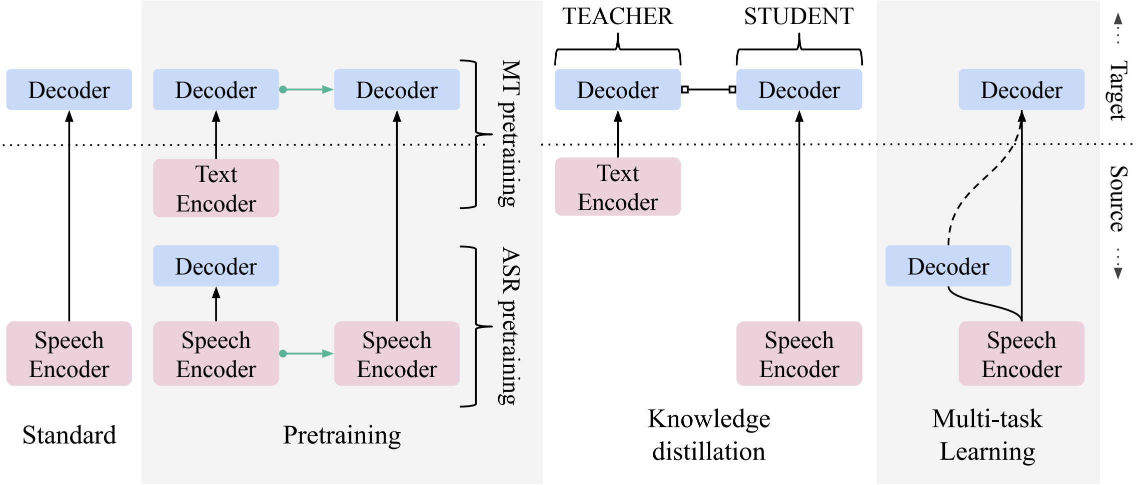
Benefits:

- Uncertainty directly handled.
  - Target-language considerations influence speech recognition.
- Potentially fewer NN parameters.

Drawbacks:

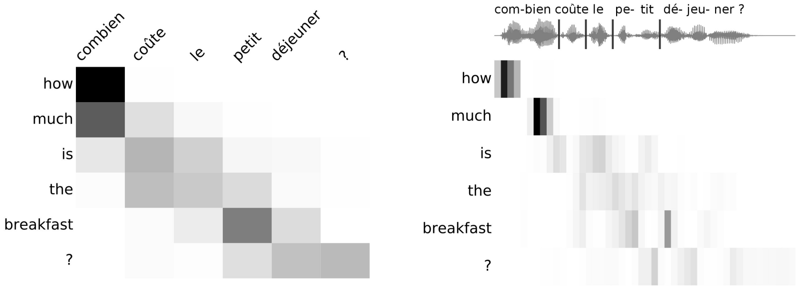
- Insufficient training data.
  - Speech + transcript and parallel texts much more common than speech + translation.
- 20–40x longer input sequences (sound timeframes vs. subwords).
- Difficult alignment problem within sentences/utterances.
- Non-golden utterance segmentation not yet much considered.

SLT Training Techniques



Proof-of-Concept End-to-End SLT (Berard et al., 2016)

- Synthetic French speech into English text (7 concatenative voices).
- MFCCs → deep LSTM encoder → attn → deep LSTM decoder.



(a) Machine translation alignment (b) Speech translation alignment

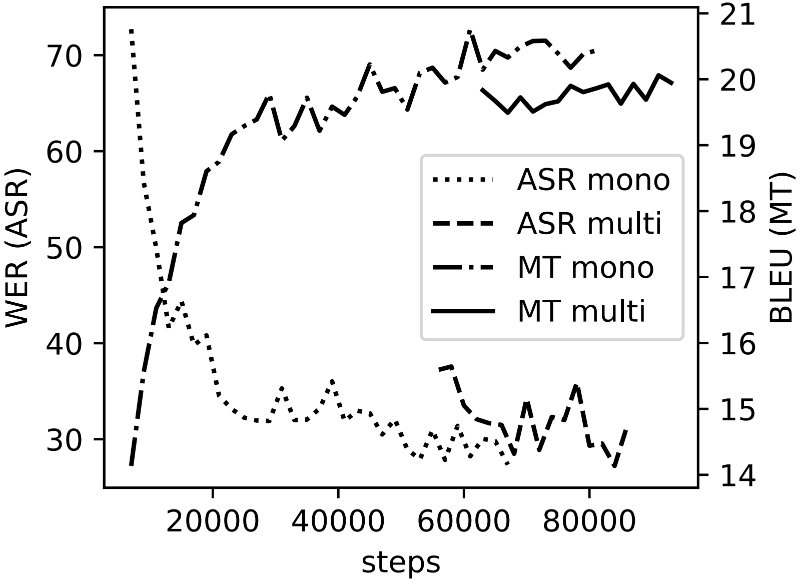
- End-to-end results not far from ASR+MT *given synthetic input*.

First Truly End-to-End SLT

Bérard et al. (2018) presents the first truly end-to-end SLT:

- Speech encoder:
    - 2 layers converting n-dim input into n'-dim.
    - 2 layers of convolution
    - 3-layer bidirectional LSTM
- Attention
- Char-level decoder
    - Used either to predict English transcription ($|V| = 46$),
    - or French translation ($|V| = 167$)

Bérard et al. (2018) Pre-Training



Bérard et al. (2018) Results

	greedy	beam	ensemble	params (million)
	Test BLEU			
Cascaded	14.6	14.6	15.8	6.3 + 15.9
End-to-End	12.3	12.9	15.5†	9.4
Pre-trained	12.6	13.3		
Multi-task	12.6	13.4		

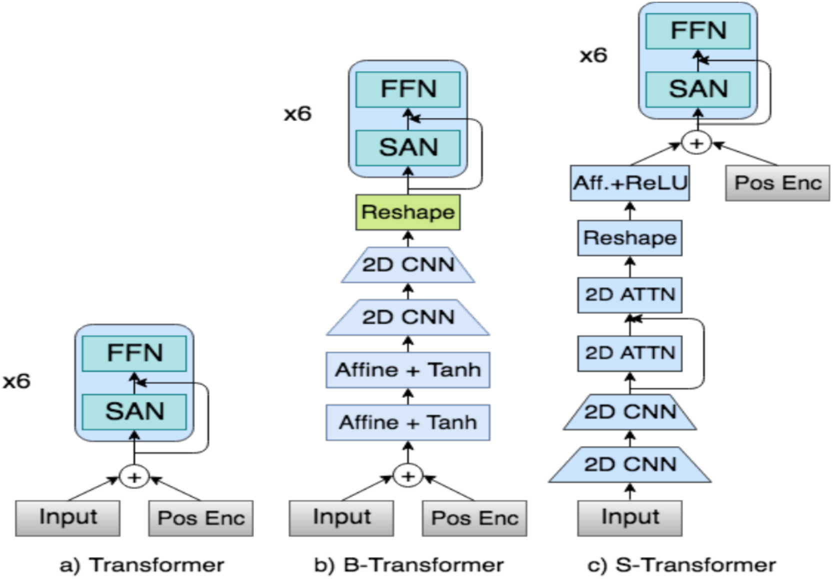
Table 4: AST results on Augmented LibriSpeech test. † combines the end-to-end, pre-trained and multi-task models.

Recent End-to-End SLT Results (Sulubacak et al., 2019)

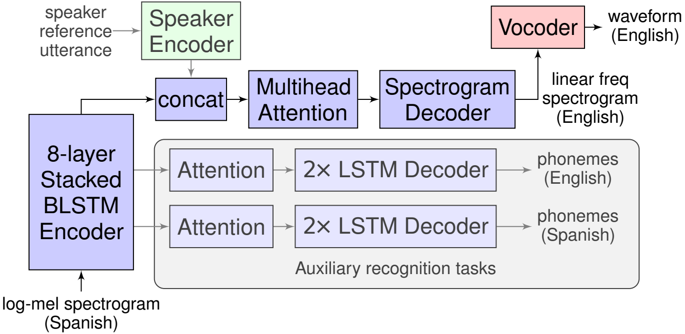
Table 4: BLEU scores for SLT methods on English→French Augmented LibriSpeech/test. All systems are end-to-end, except for the pipeline system marked with a dagger (†).

Approach	BLEU ↑	Training data			Description
		SLT (h)	ASR (h)	MT (sent)	
Bérard et al (2018)	13.4	100h			CNN+LSTM. Multi-task.
Di Gangi et al (2019b)	13.8	236h			CNN+Transformer.
Bahar et al (2019)	17.0	100h	130h	95k	Pyramidal LSTM. Pretraining, augmentation.
Liu et al (2019)	17.0	100h			Transformer. Knowledge distillation.
Inaguma et al (2019a)	17.3	472h			CNN+LSTM. Multilingual.
Pino et al (2019)	21.7	100h	902h	29M	CNN+Transformer. Pretraining, augmentation.
Pino et al (2019) †	21.8	100h	902h	29M	End-to-end ASR. CNN+LSTM.

Transformer Adapted for Speech Input Gangi et al. (2019)



Translatotron (Jia et al., 2019)



- Speech transcripts still needed to train (but not at inference).
- Somewhat worse that SLT+TTS.
- Allows to transfer the voice across languages.

`https://google-research.github.io/lingvo-lab/translatotron/`

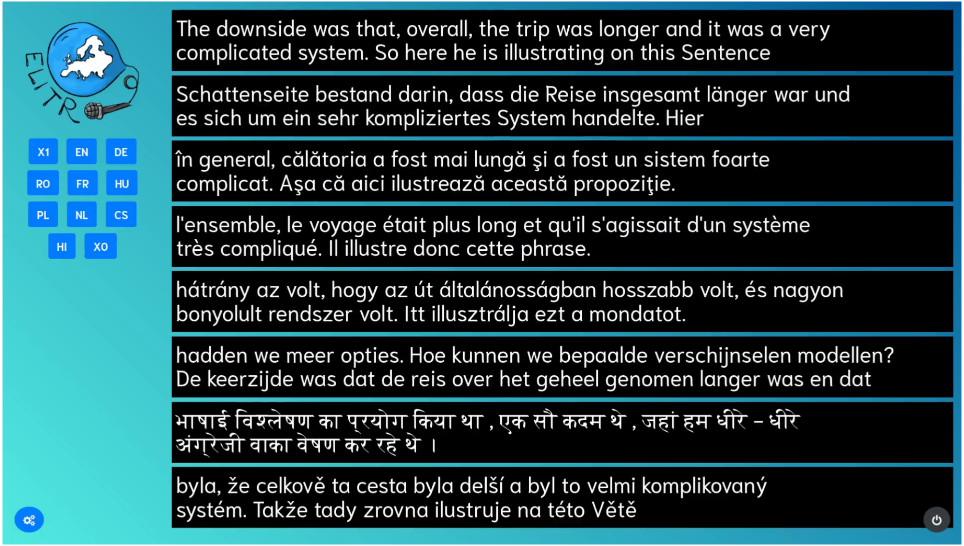
Spoken Language Translation

Presentation

The Importance of Presentation

- Presentation issues can kill the whole show.
- Bad font size may make output impossible to follow.
- Too much flicker, jumping text, ...
    - Recent fully NN ASR operate on a moving window of say 8 seconds.
    - The output is too unstable to follow, let alone if translated by MT.
- Presentation must be tested on stage.
    - Sizing, visibility, ... cannot be checked remotely.

Subtitle View



Paragraph View

evaluation so that we have the English sentences original from some English newspapers and the reference translations we get made by some professional translator, that's what they're saying.

27. Like the referendum is actually a translation, not the original sentence.
28. Well, of course, that's already here.
29. The problem I mentioned.
30. The translation may not be adequate.
31. Unless the translator did it the right way.
32. Or it may not be fully fluid.
33. And how was the first slide?
34. Occasionally the translator accent One mistake or another mistake Occasionally a different approach is used, when I take originally Czech sentences z.
35. Let's say from originally Czech newspapers

die englischen Urteile haben, die aus einigen englischen Zeitungen stammen, und die Referenzübersetzt uns von einem Berufsübersetzenden.

27. Wie bei dem Referendum handelt es sich eigentlich um eine Übersetzung, nicht um den ursprünglichen Satz.
28. Ja, das ist ja schon hier.
29. Das Problem, das ich erwähnt habe.
30. Vielleicht ist die Übersetzung nicht ausreichend.
31. Es sei denn, der Übersetzer hat es richtig gemacht.
32. Oder es ist vielleicht nicht völlig flüssig.
33. Und wie war der erste Verfall?
34. Gelegentlich wird bei einem Übersetzungsfehler oder einem anderen Fehler eine andere Vorgehensweise angewandt, wenn ich ursprünglich tschechische Urteile z nehme.
35. Sagen wir aus den ursprünglichen tschechischen Zeitungen.

že máme anglické věty původní z nějakých anglických novin a referenční překlady si necháme vyrobit nějakým profesionálním překladatelem přeložit, tak to tady označují.

27. Jako že ta referend Je ve skutečnosti překladem, nikoli tou původní větou.
28. Tak samozřejmě, to už je tady.
29. Ten problém, který jsem zmiňoval.
30. Ten překlad nemusí být adekvátní.
31. Pokud to ten překladatel neudělal úplně správně.
32. Nebo nemusí být plně plynulý.
33. A jak bylo na tom prvním slidu?
34. Občas ten překladatel přízvuk Jednu chybu nebo druhou chybu Občas se používá jiný přístup, kdy vezmu původně české věty z.
35. Dejme tomu z původně českých novin

EN
DE
CS
HI
X0
X1

ASR/Segmentation Updates (Cho et al., 2012; Cho et al., 2017)

stable ASR output *unstable ASR output* *ASR update which arrives later*

ASR pixels on your screen at any given moment it is also very flexible architecture of this is an entire book

stable segmenter output *unstable segmenter output*

SEG1 pixels on your screen. At any given moment. It is also very flexible architecture of...

completed *expected* *incoming*

stable segmenter output *unstable segmenter output*

SEG2 pixels on your screen. At any given moment. It is also very flexible. Architecture of this is an entire book.

completed *completed* *expected* *expected*

MT1 ...Pixelen auf Ihrem Bildschirm. Zu jedem Zeitpunkt. Es ist auch eine sehr flexible Architektur...

MT2 ...Pixelen auf Ihrem Bildschirm. Zu jedem Zeitpunkt. Sie ist auch sehr flexibel. Die architektur ist ein ganzes Buch.

should reset happen, this would be the new beginning

Cognitive Load, Overall Usability

- Users confirm that transcript and slides must be on the same screen.
  - Adding slide streaming/sharing to both Subtitle and Paragraph view.
- Overall usability:
  - Often still bad, due to the cummulation of errors.
  - Two foreign colleagues reported they could follow a Czech talk, if fully focussed on the text.
- Desired settings differ from user to user:
  - Those who understand source language will need simultaneity over precision and stability.
  - Those who cannot understand source need stability and precision and are happy to wait for *seconds*.

Spoken Language Translation

Evaluation

Evaluating Spoken Language Translation

Three aspects o simultaneous (‘on-line’) SLT:

- Quality of the final translation.
  - … equals standard MT quality estimates.
- Lag behind the source.
  - Some lag is inevitable, e.g. waiting for the German verb.
- Flicker
  - How many words are corrected?

Mismatch in Segmenting

English ASR	...	Do you know it's a cat,		Isn't it? yes is it like that.	...
German Ref	...	Wisst ihr es?	Es ist eine Katze, ist es das nicht?	Ja, so ist es.	...

- Consider English→German SLT.
- No matter what the MT does with the recognized English, segments won't match.

Mismatch in Segmenting

English ASR	...	Do you know it's a cat,	Isn't it? yes is it like that.	...

German Ref	...	Wisst ihr es?	Es ist eine Katze, ist es das nicht?	Ja, so ist es.	...

Planned strategy:

- Follow reference segmentation.
- Find best matching hypothesis segmentation.
  - a) Expand by full segments.
  - b) Expand by a few words around the best-matching segment.

Or ignore the problem by force-segmenting into $\sim$30s chunks.

Visual Information in MT

Motivation for Multi-Modal Translation (1/2)

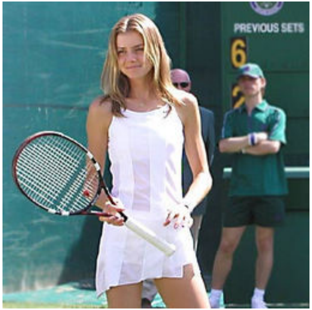
Input	A tennis player is getting ready.
Output A	Tenista se připravuje.

Motivation for Multi-Modal Translation (1/2)

Input	A tennis player is getting ready.
Output A	Tenista se připravuje. $\leftarrow$ male
Output B	Tenistka se připravuje. $\leftarrow$ female

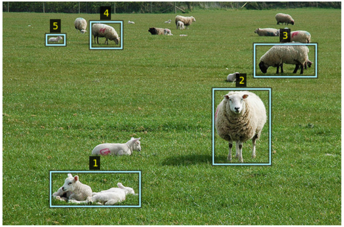
Motivation for Multi-Modal Translation (1/2)

Input	A tennis player is getting ready.
Output A	Tenista se připravuje. ← male
Output B	Tenistka se připravuje. ← female



Motivation for Multi-Modal Translation (2/2)

Hindi Visual Genome (Parida et al., 2019) provides 30k picture descriptions from `visualgenome.org`, translated into Hindi.



1: Two lambs lying in the sun.

Hindi MT: दो भेड़ के बच्चे सूरज में **झूठ बोल** रहे हैं

Gloss: Two baby sheep are **telling lies** ...

Selected surrounding captions:

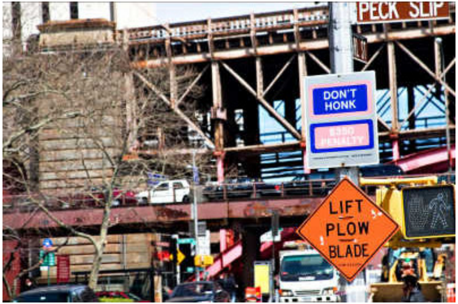
2. Sheep standing in the grass
3. Sheep with black face and legs
4. Sheep eating grass
5. Lamb sitting in grass.

Hindi Visual Genome Challenge Test Set

- A test set created by scanning the 3.15M unique strings for ambiguous words.
- Only 19 words with multiple (automatic) translations were identified:

	Word	Segment Count		Word	Segment Count
1	Stand	180	11	English	42
2	Court	179	12	Fair	41
3	Players	137	13	Fine	45
4	Cross	137	14	Press	35
5	Second	117	15	Forms	44
6	Block	116	16	Springs	30
7	Fast	73	17	Models	25
8	Date	56	18	Forces	9
9	Characters	70	19	Penalty	4
10	Stamp	60		Total	1400

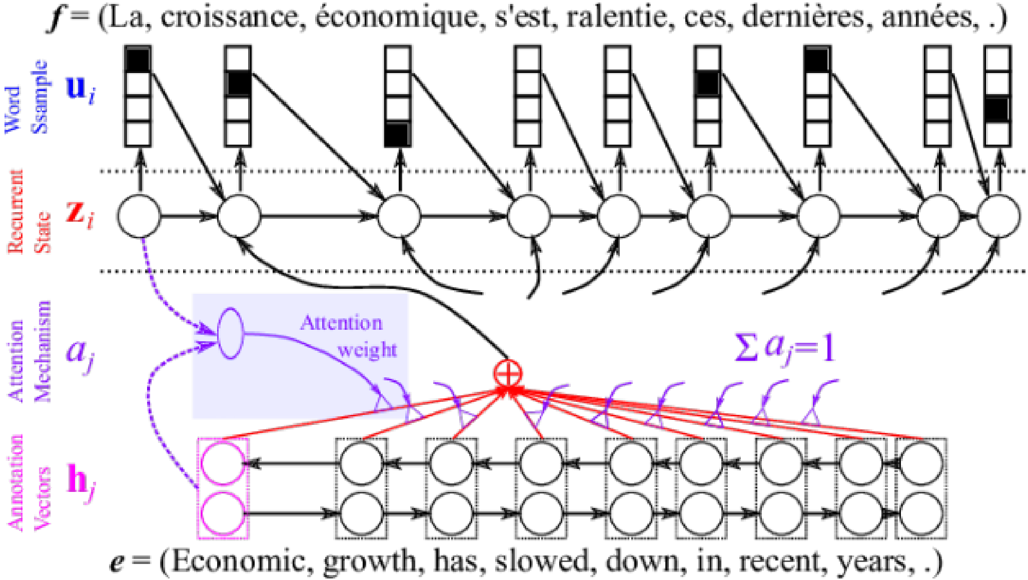
Example from the "Challenge Test Set"



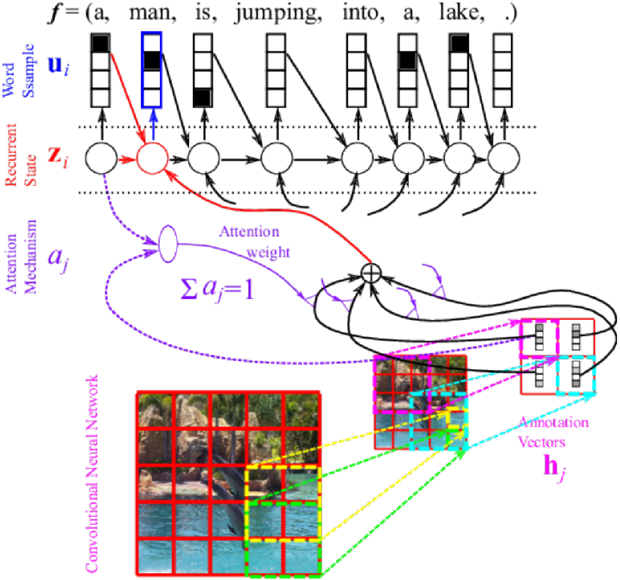
Street sign advising of penalty.

The penalty box is white lined.

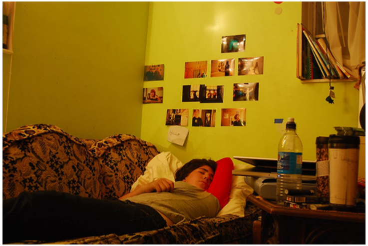
Attention to Source Words (?)



Attention to Source Image



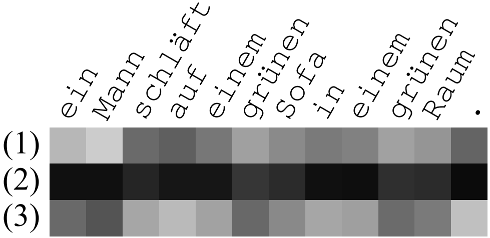
Hierarchical Attention (Libovický and Helcl, 2017)



Source: a man sleeping in a green room on a couch .

Reference: ein Mann schläft in einem grünen Raum auf einem Sofa .

Output with attention:



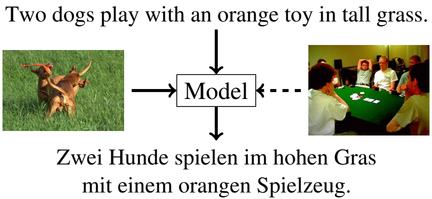
(1) source, (2) image, (3) sentinel

Recent Multi-Modal MT Results (Sulubacak et al., 2019)

	BLEU ↑	METEOR ↑	Type	Description	Arch.
Elliott et al (2015) †	9.7 (N/A)	24.7 (N/A)	E,D	Conditional LMs	RNN
Caglayan et al (2016a) †	29.3 (↓4.6)	48.5 (↓4.3)	A	Shared Attention	RNN
Calixto et al (2016) †	28.8 (N/A)	49.6 (N/A)	A	Separate Attention	RNN
Huang et al (2016) †	36.8 (↑2.0)	54.4 (↑2.3)	IF	Parallel RCNN-LSTMs	RNN
Hitschler et al (2016) †	34.3 (N/A)	56.0 (N/A)	R	Retrieval + Reranking	SMT
Toyama et al (2016)	36.5 (↑1.6)	56.0 (↑0.7)	L	Variational	RNN
Shah et al (2016) †	34.8 (↑0.2)	56.7 (↑0.1)	R	Visual Reranking	SMT
Caglayan et al (2016a) †	36.2 (- 0.0)	57.5 (↑0.1)	R	Visual Reranking	SMT
Helcl and Libovicky (2017)	31.9 (↓2.7)	49.4 (↓2.3)	A	Hierarchical Attention	RNN
Calixto and Liu (2017)	36.9 (↑3.2)	54.3 (↑2.0)	I	Input Prepend & Append	RNN
Calixto et al (2017)	36.5 (↑2.8)	55.0 (↑2.7)	A	Gated Attention	RNN
Calixto and Liu (2017)	37.3 (↑3.6)	55.1 (↑2.8)	D	Decoder Init.	RNN
Elliott and Kadar (2017)	36.8 (↑1.3)	55.8 (↑1.8)	T	Imagination	RNN
Caglayan et al (2017a)	38.2 (↑0.1)	57.6 (↑0.3)	E,D	Encoder Decoder Init.	RNN
	37.8 (↓0.3)	57.7 (↑0.4)	O	Multiplicative Interaction	RNN
Delbrouck and Dupont (2017b)	40.5 (N/A)	57.9 (N/A)	A	Encoder Attention + CBN	RNN
Arslan et al (2018)	41.0 (↑2.4)	53.5 (↓1.5)	A	Parallel Attention	Transformer
Calixto et al (2018)	37.6 (↑2.6)	56.0 (↑1.1)	L	Variational	RNN
Helcl et al (2018b)	38.8 (↑0.7)	56.4 (↑0.2)	T	Imagination	Transformer
Libovicky et al (2018)	38.5 (↑0.2)	56.5 (↓0.2)	A	Hierarchical Attention	Transformer
	38.6 (↑0.3)	57.4 (↑0.7)	A	Parallel Attention	Transformer
Ive et al (2019)	38.0 (↑0.1)	55.6 (↓0.3)	DF	2-stage Decoder + Label Embs.	Transformer
Libovicky (2019)	37.6 (↑0.9)	56.0 (↑0.9)	A	Hierarchical Attention	RNN
Caglayan (2019)	39.0 (↑0.1)	58.5 (↑0.1)	E,D	Encoder Decoder Init.	RNN
	39.4 (↑0.5)	58.7 (↑0.3)	A	Separate Attention + L_2 Norm.	RNN
Unconstrained ensembles					
Helcl et al (2018b)	42.6 (↑2.2)	59.4 (↑0.4)	T	Imagination	Transformer
Grönroos et al (2018)	45.5 (- 0.0)	(N/A)	IF	Input Prepend	Transformer

Is Visual Information Needed? (1/2)

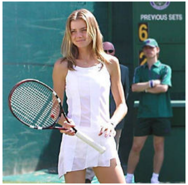
- Our text-only English→Hindi was perfect on the “challenge” words.
- Elliott (2018) used MM systems with shuffled, incongruent images.



  - Only the hierarchical attention was sensitive to images
    other multi-modal systems performed equally with congruent and incongruent images.
- Caglayan et al. (2019) list other papers where images have not helped much.

Is Visual Information Needed? (2/2)

Elliott (2018) degrade the textual input
and show that multi-modal MT performs better:



SRC : a young [v] in [v] holding a tennis [v]
NMT : un jeune garçon en bleu tenant une raquette de tennis
(a young boy in blue holding a tennis racket)
MMT : une jeune femme en blanc tenant une raquette de tennis
REF : une jeune femme en blanc tenant une raquette de tennis
(a young girl in white holding a tennis racket)

SRC : little girl covering her face with a [v] towel
NMT : une petite f lle couvrant son visage avec une serviette blanche
(a little girl covering her face with a white towel)
MMT : une petite f lle couvrant son visage avec une serviette bleue
REF : une petite f lle couvrant son visage avec une serviette bleue
(a little girl covering her face with a blue towel)

Summary

- Speech translation:
  - Simple cascading suffers from uncertainty loss, error cummulation.
  - Problems with segmentation.
  - End-to-end systems recently approaching cascaded ones.
  - Practical deployment of live subtitling is a challenge.
- Translation with visual features:
  - Motivation: Image can be the missing context for ambiguity resolution.
  - Discussion on image utility.

A picture is worth a thousand words

Summary

- Speech translation:
  - Simple cascading suffers from uncertainty loss, error cummulation.
  - Problems with segmentation.
  - End-to-end systems recently approaching cascaded ones.
  - Practical deployment of live subtitling is a challenge.
- Translation with visual features:
  - Motivation: Image can be the missing context for ambiguity resolution.
  - Discussion on image utility.

A picture is worth a thousand words
in one of a thousand cases.

References

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. CoRR, abs/1612.01744. Published at NIPS.

A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6224–6228.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4159–4170, Minneapolis, Minnesota, June. Association for Computational Linguistics.

E. Cho, J. Niehues, and A. Waibel. 2012. Segmentation and punctuation prediction in speech language translation using a monolingual translation system. In Proceedings of the Ninth International Workshop on Spoken Language Translation (IWSLT).

Eunah Cho, Jan Niehues, and Alex Waibel. 2017. NMT-based segmentation and punctuation insertion for real-time spoken language translation. In Interspeech 2017. ISCA, aug.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2019. Adapting Transformer to End-to-End Spoken Language Translation. In Proc. Interspeech 2019, pages 1133–1137.

Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. CoRR, abs/1904.06037.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In