Overview of the IWSLT 2017 Evaluation Campaign

Tokyo, December 14-15

Summary

- Mission of IWSLT
- This year evaluation
  - [Detailed reports by colleagues]
- Reflections and outlook

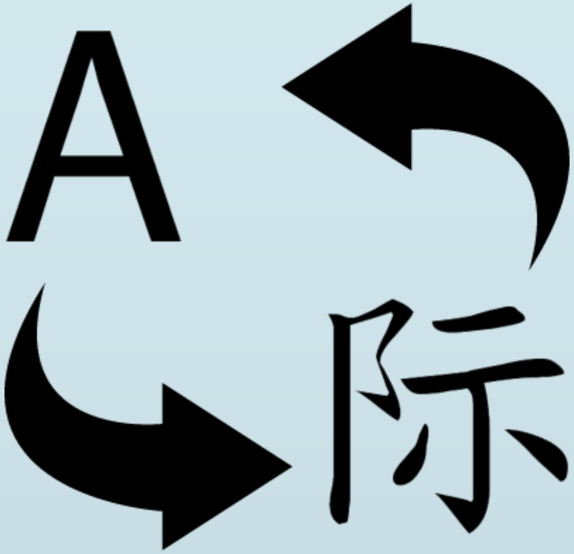
Mission

- Spoken language translation
- Evaluation framework
- Challenging tasks

Mission

- Spoken Language Translation

Speech Recognition

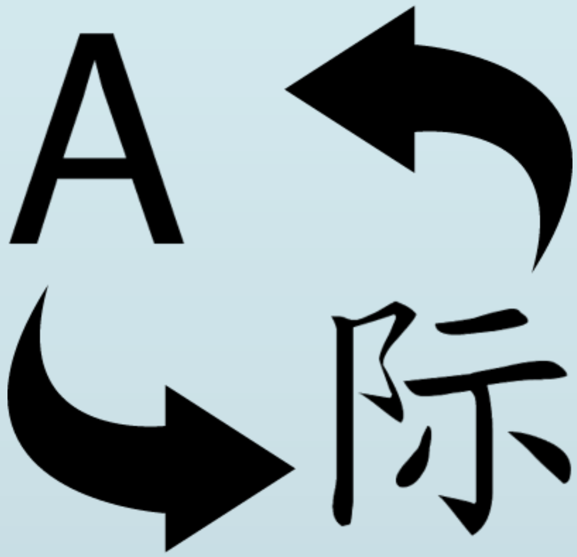


Machine Translation

Mission

- Spoken Language Translation

“And he emailed me this picture.”



Machine
Translation

Mission

- Spoken Language Translation

“an emailed me this picture”

Speech
Recognition

Mission

Mission

- Challenging tasks

Mission

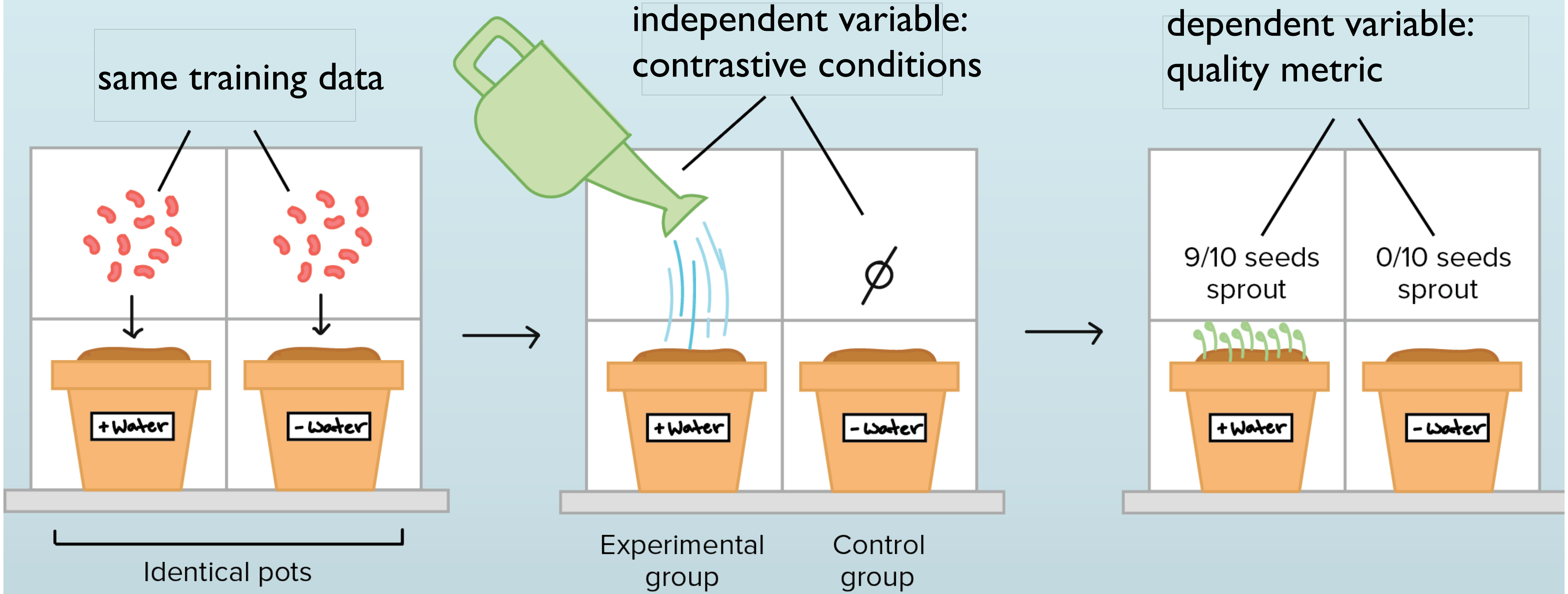
- Evaluation framework

Mission

- Evaluation framework

Mission

- Evaluation framework



same training data

independent variable:
contrastive conditions

dependent variable:
quality metric

Identical pots

Experimental group

Control group

9/10 seeds sprout

0/10 seeds sprout

Mission

- Challenging tasks: traveling domain (from 2004)

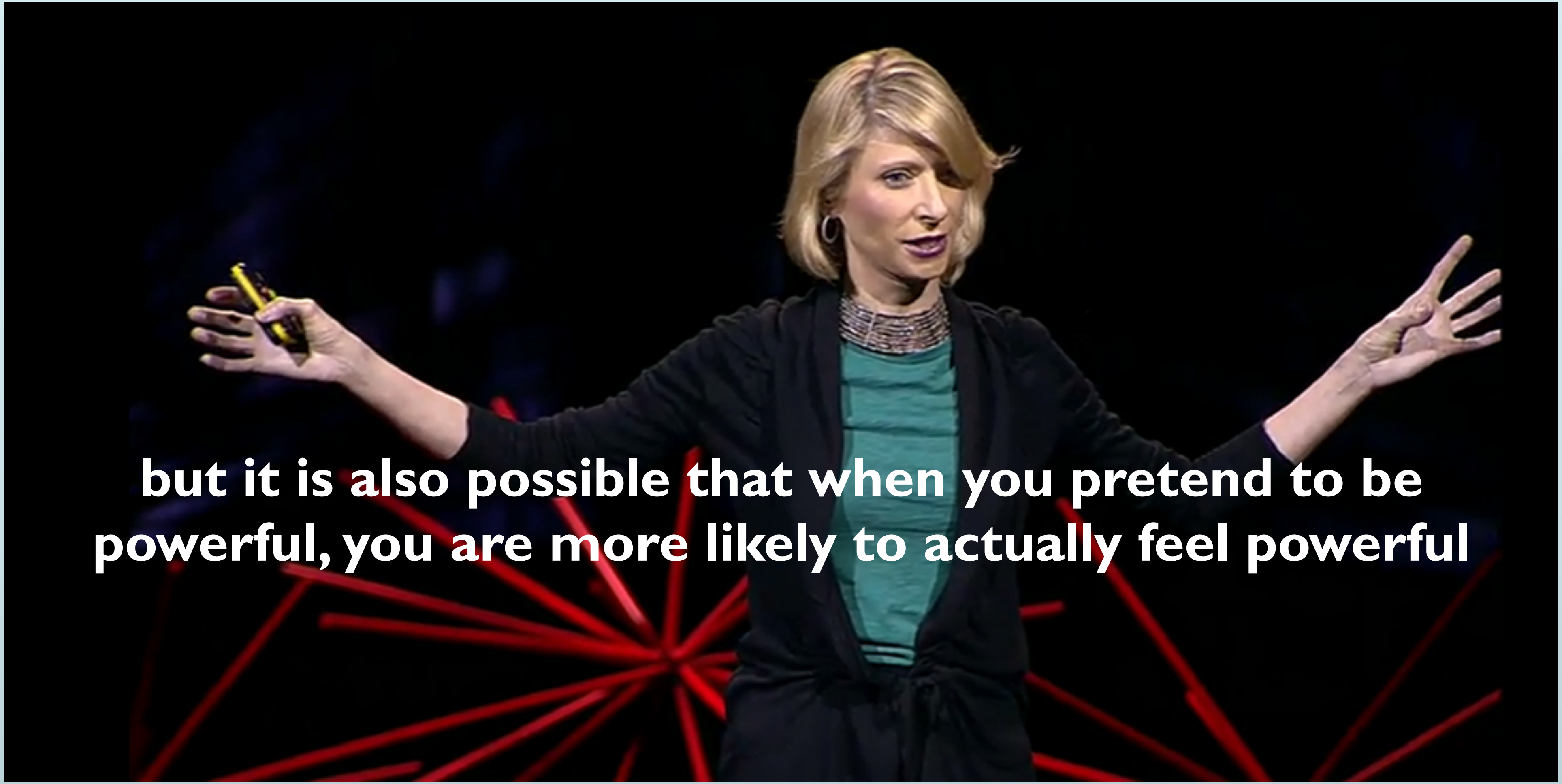
Mission

- Challenging tasks: traveling domain (from 2004)

"I WANT A FAMILY OF 5, INTERCONNECTING CABINS ON THE CRUISE DEPARTING TONIGHT."

Mission

- Challenging tasks: TED talks (from 2011)



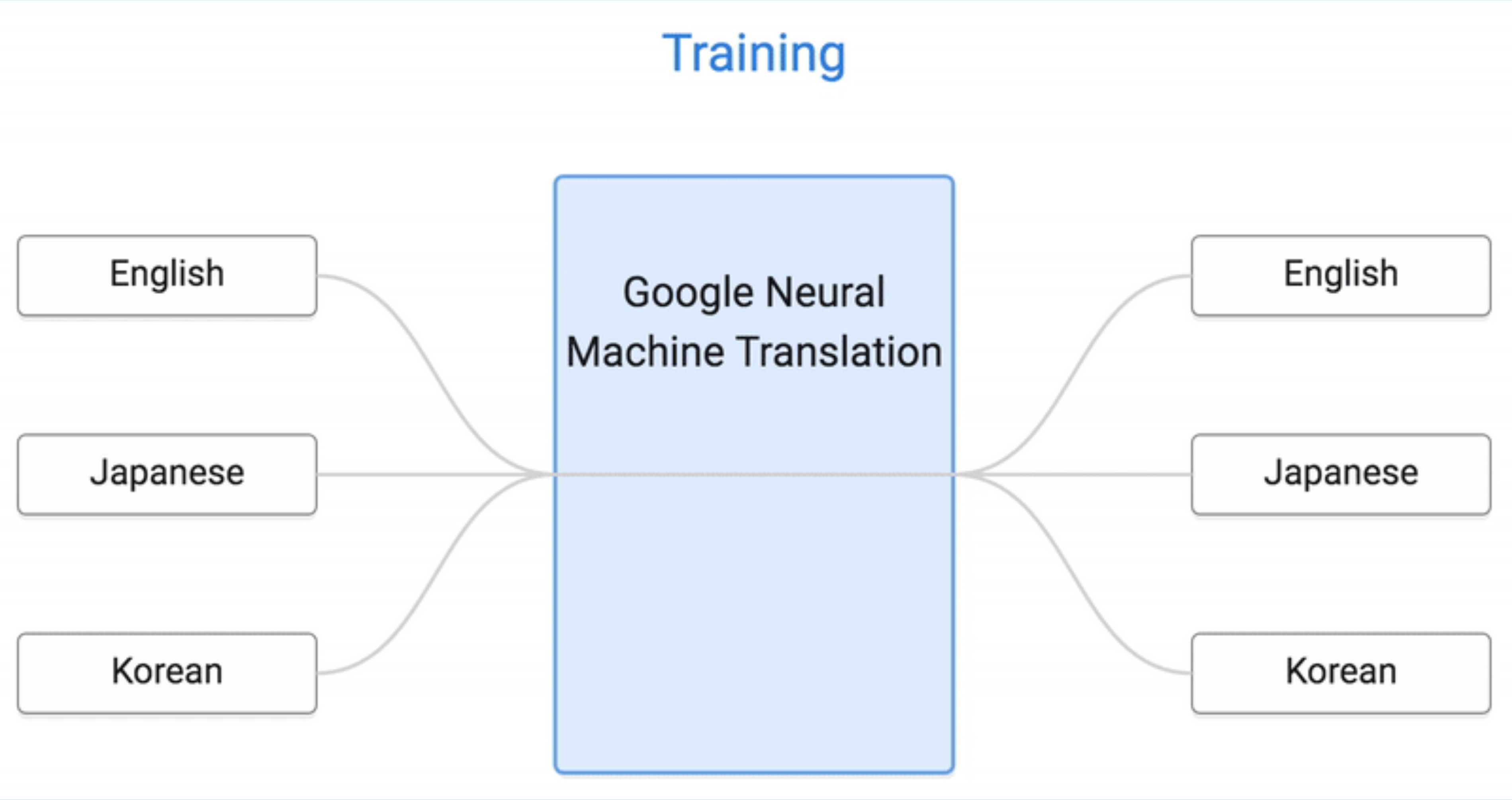
but it is also possible that when you pretend to be powerful, you are more likely to actually feel powerful

This year evaluation

- Multilingual task (TED Talks)
- Dialogue task
- Lecture task

This year evaluation

- Multilingual task



This year evaluation

- Dialogue task

D: How is your back pain today?

This year evaluation

- Dialogue task

D: How is your back pain today?

P: I don't have any, I never actually got any.

This year evaluation

- Dialogue task

D: How is your back pain today?
P: I don’t have any, I never actually got any.

D: Wie sind Ihre Rückenschmerzen heute?
P: I habe keine, Ich habe eigentlich nie welche bekommen.

This year evaluation

- Lecture task

well the reason why … neural machine translation output is so good … well … I don't know … actually nobody knows!

This year evaluation

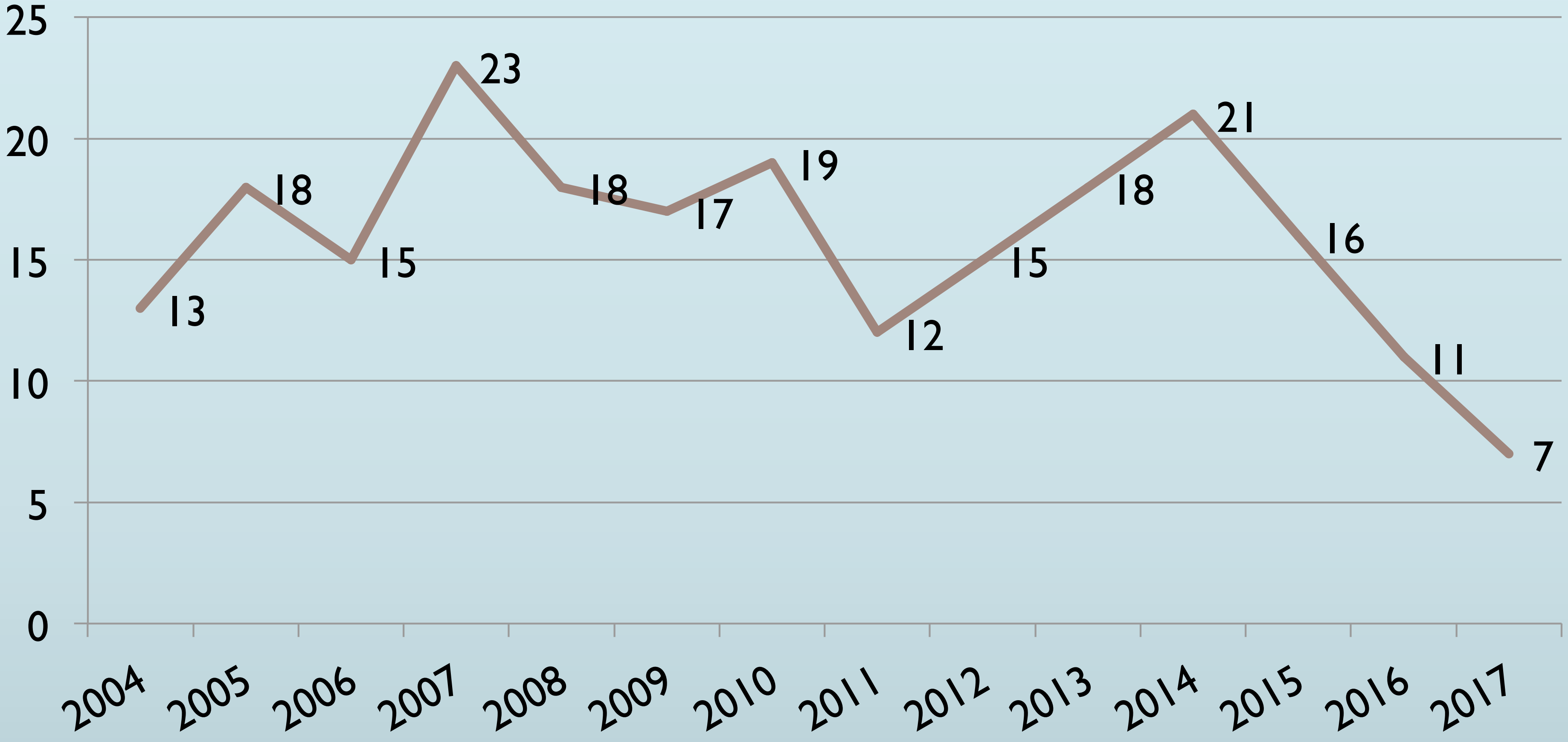
- Multilingual task – report by Luisa Bentivogli
- Lecture task – report by Jan Niehues
- Dialogue task – report by Katsuhito Sudoh

Reflections and outlook

- Drop of evaluation participants
- Increasing interest in the benchmark
- Discussion

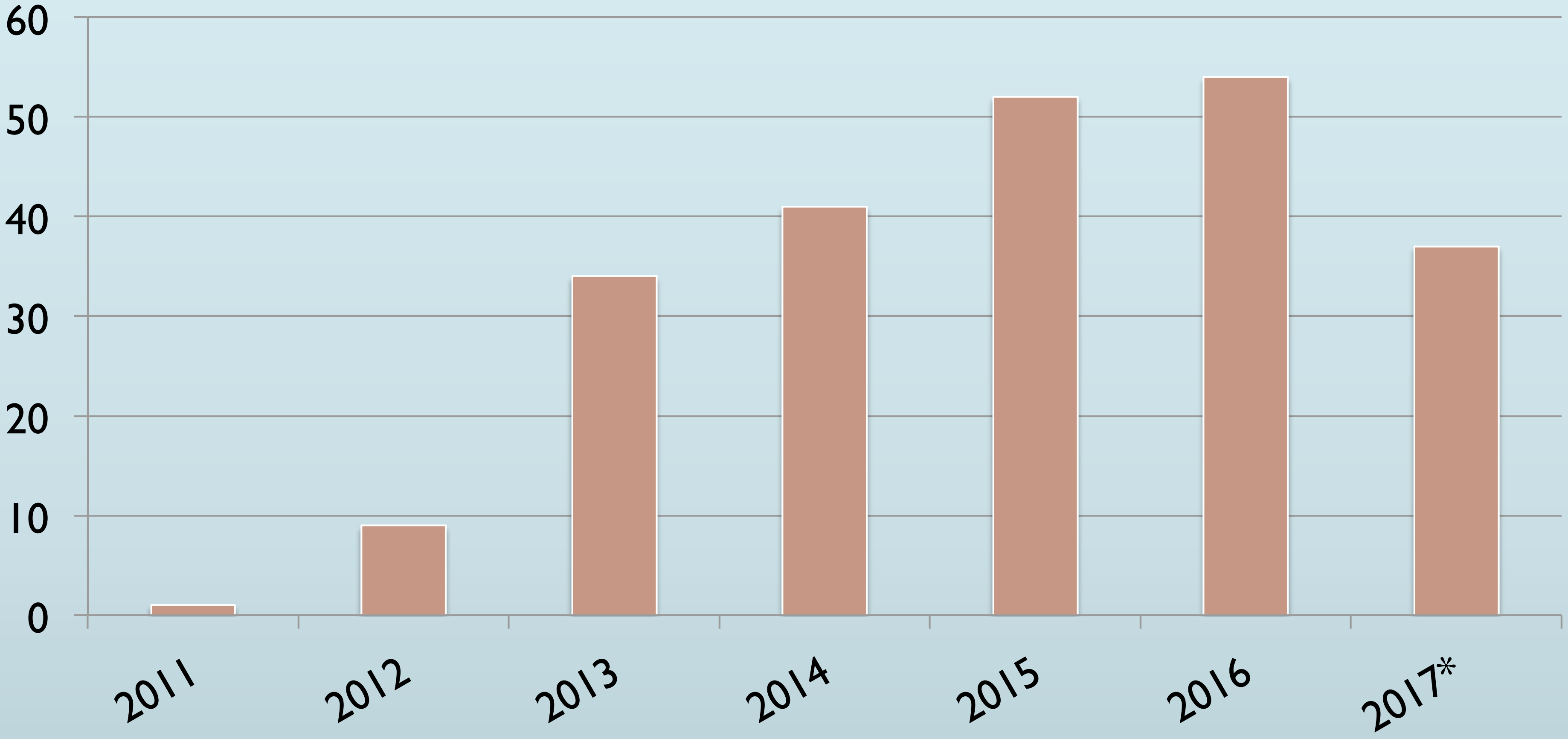
Reflections and outlook

- Drop in participation



Reflections and outlook

- Citations of the TED Talk benchmark paper



Reflections and outlook

- Discussion

Possible issues:

- Participation model of WMT seems more successful
- Two evaluations in a year are maybe too much
- Lack of interest in the speech side of SLT
  - people look at IWSLT as another MT evaluation
- IWSLT as a standalone event is less attractive
- Timing of IWSLT often overlaps with other events

Reflections and outlook

- A few options

- Try to move some IWSLT evaluation tasks to WMT
- Co-locate IWLST with some other conferences
  - ACL group? ACL, EACL, NAACL, EMNLP
  - IMTA group? MT Summit, AMTA, EAMT
- Promote IWSLT benchmarks instead of evaluations
  - people can run tests whenever they want and present their results at the workshop
- We would like to collect feedback from the participants

goo.gl/XYC2hZ

Question time

IWSLT 2017

Multilingual Task
Human Evaluation

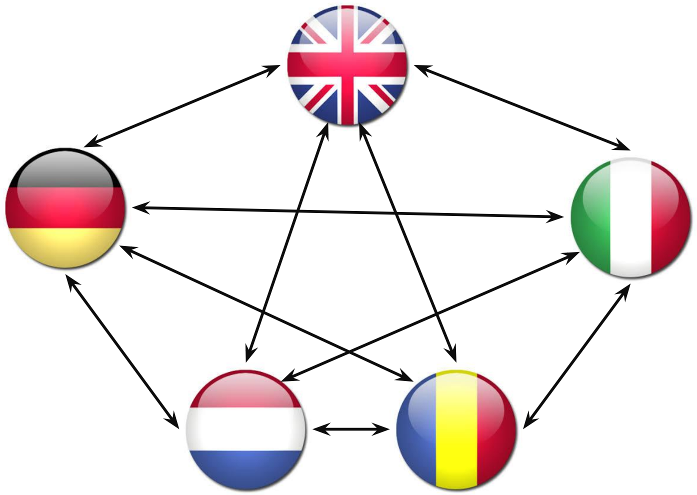
Luisa Bentivogli[1], Christian Federmann[2]

[1]Fondazione Bruno Kessler, Trento, Italy
[2]Microsoft AI+Research - Redmond, WA, USA

Multilingual Task

1 MT system for

20 language directions



Training data conditions:

- **Large Data -** long list of permissible resources
- **Small Data -** in-domain data only
  - average for each direction:

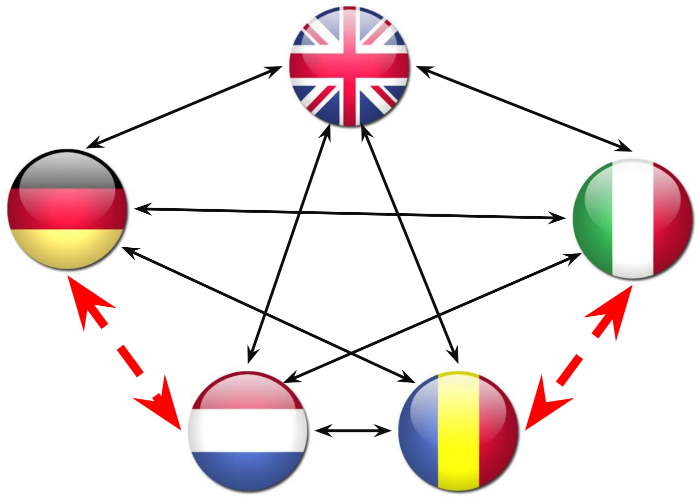
    1749 TED talks, ~200k sentences, ~4M tokens

Multilingual Task: Zero-Shot Translation

Tested on 4 language directions

Dutch <-> German

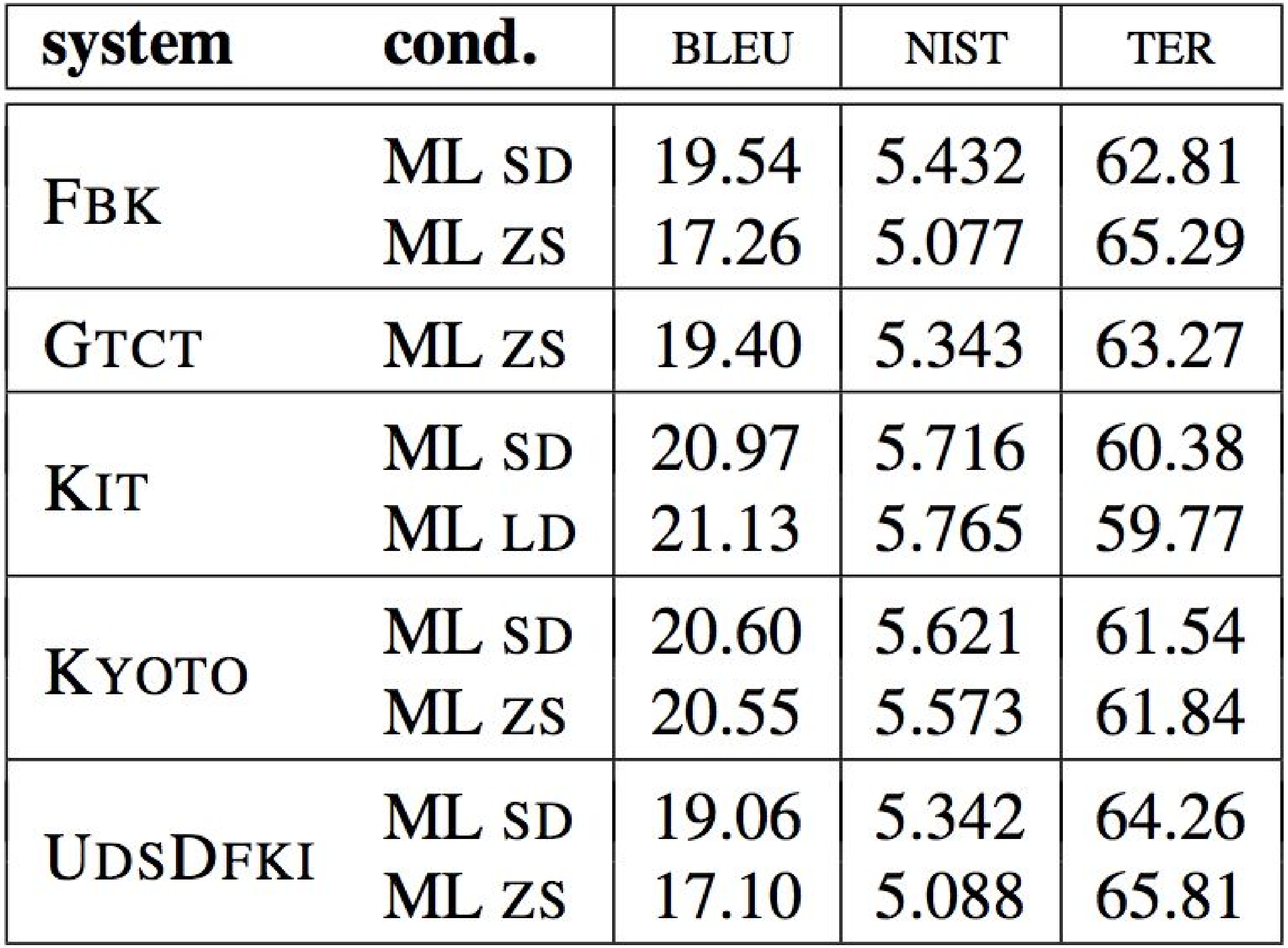
Romanian <-> Italian



- ***No training data*** for the 4 tested directions
- ***Small data*** condition for the other 16 language directions in the multilingual system

Automatic Evaluation

Average results for the 4 zero-shot directions:



system	cond.	BLEU	NIST	TER
FBK	ML SD	19.54	5.432	62.81
	ML ZS	17.26	5.077	65.29
GTCT	ML ZS	19.40	5.343	63.27
KIT	ML SD	20.97	5.716	60.38
	ML LD	21.13	5.765	59.77
KYOTO	ML SD	20.60	5.621	61.54
	ML ZS	20.55	5.573	61.84
UDSDFKI	ML SD	19.06	5.342	64.26
	ML ZS	17.10	5.088	65.81

Human Evaluation

Focus:

- Zero-Shot Translation Task

Additional systems:

- Bilingual (*small data*) for *Nl-De* and *Ro-It*

HE dataset:

- Subset of *tst2017* **-** 10 TED Talks

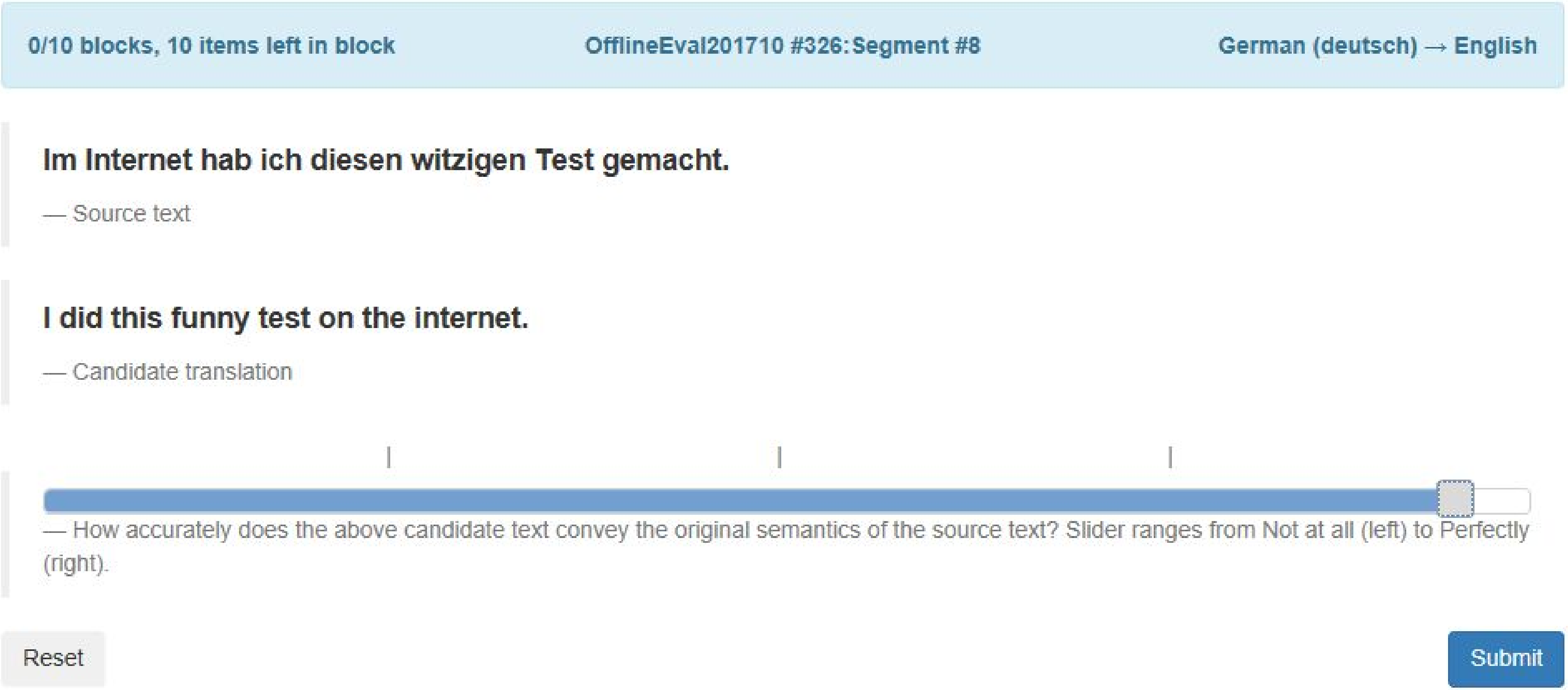
HE methodologies:

- Direct Assessment - official ranking (*Nl <-> De, Ro <-> It*)
- Post-Editing - comparison of ML - ZS / BL (*Nl->De, Ro->It*)

Direct Assessment

Assessment of the overall MT translation quality based on the accuracy wrt

- Source sentence
- Reference translation



0/10 blocks, 10 items left in block OfflineEval201710 #326:Segment #8 German (deutsch) → English

Im Internet hab ich diesen witzigen Test gemacht.

— Source text

I did this funny test on the internet.

— Candidate translation

— How accurately does the above candidate text convey the original semantics of the source text? Slider ranges from Not at all (left) to Perfectly (right).

Direct Assessment Setup

- DA scores for 300 sentences (half the HE dataset)
- Double redundancy for all data points collected
- Annotation done by trained linguistic consultants
- Using Appraise evaluation framework (same as for WMT17)

Language	**Annotators**	**Tasks**	**Redundancy**	**Tasks/ annotator**	**Total tasks**
Dutch→German	a=22	t=55	r=2	5	110
Romanian→Italian	a=22	t=55	r=2	5	110
German→Dutch	a=16	t=40	r=2	5	80
Italian→Romanian	a=16	t=40	r=2	5	80

Results: Dutch→German

#	Ave %	Ave z	System	Condition
1	70.2	0.173	KIT	ML LD
2	70.2	0.145	Kyoto	BL SD
	69.4	0.139	Kyoto	ML SD
3	68.1	0.110	KIT	ML SD
4	68.4	0.103	Kyoto	ML ZS
	66.5	0.040	GTCT	ML ZS
	67.0	0.029	UDS-DFKI	ML SD
5	64.5	-0.045	FBK	BL SD
	63.5	-0.078	UDS-DFKI	ML ZS
	63.3	-0.079	FBK	ML SD
6	60.0	-0.212	FBK	ML ZS
7	57.2	-0.338	UDS-DFKI	BL SD

Source-based DA

#	Ave %	Ave z	System	Condition
1	64.2	0.121	KIT	ML LD
2	63.5	0.100	Kyoto	ML SD
3	64.6	0.102	Kyoto	ML SD
4	63.0	0.069	Kyoto	ML ZS
	62.1	0.061	KIT	ML SD
	62.7	0.045	UDS-DFKI	ML SD
	61.2	0.014	GTCT	ML ZS
5	61.1	0.017	FBK	BL SD
6	59.2	-0.076	UDS-DFKI	ML ZS
	58.0	-0.092	FBK	ML SD
7	56.2	-0.178	FBK	ML ZS
	54.9	-0.241	UDS-DFKI	BL SD

Reference-based DA

Results: Romanian→Italian

#	Ave %	Ave z	System	Condition
1	74.8	0.222	Kyoto	BL SD
2	74.4	0.200	KIT	ML SD
	72.1	0.131	Kyoto	ML SD
3	72.1	0.136	Kyoto	ML ZS
	71.8	0.115	KIT	ML LD
4	71.1	0.081	UDS-DFKI	ML SD
	70.3	0.049	FBK	ML SD
	69.1	0.017	GTCT	ML ZS
	68.5	0.000	FBK	BL SD
5	66.9	-0.090	UDS-DFKI	ML ZS
6	61.6	-0.268	FBK	ML ZS
7	55.3	-.0546	UDS-DFKI	BL SD

Source-based DA

#	Ave %	Ave z	System	Condition
1	59.9	0.169	KIT	ML SD
2	59.9	0.162	Kyoto	ML SD
3	58.9	0.126	Kyoto	BL SD
	58.6	0.126	Kyoto	ML ZS
	58.3	0.102	KIT	ML LD
4	58.3	0.086	UDS-DFKI	ML SD
5	55.2	0.014	GTCT	ML ZS
	55.1	-0.010	FBK	ML SD
	54.0	-0.045	FBK	BL SD
	54.0	-0.047	UDS-DFKI	ML ZS
6	49.0	-0.190	FBK	ML ZS
7	42.9	-0.423	UDS-DFKI	BL SD

Reference-based DA

Results: German→Dutch

#	Ave %	Ave z	System	Condition
1	70.3	0.128	Kyoto	ML ZS
2	70.0	0.088	KIT	ML LD
3	69.8	0.094	Kyoto	ML SD
	67.5	0.015	GTCT	ML ZS
	67.5	-0.002	KIT	ML SD
	67.4	-0.006	FBK	ML SD
4	66.5	-0.022	UDS-DFKI	ML SD
	66.0	-0.073	UDS-DFKI	ML ZS
5	62.4	-0.180	FBK	ML ZS

Source-based DA

#	Ave %	Ave z	System	Condition
1	57.7	0.126	KIT	ML LD
2	57.7	0.119	Kyoto	ML SD
	56.6	0.090	Kyoto	ML ZS
3	54.7	0.004	KIT	ML SD
4	54.4	0.009	GTCT	ML ZS
	53.7	-0.022	UDS-DFKI	ML SD
	53.4	-0.068	UDS-DFKI	ML ZS
	52.6	-0.073	FBK	ML SD
5	50.2	-0.156	FBK	ML ZS

Reference-based DA

Results: Italian→Romanian

#	Ave %	Ave z	System	Condition
1	77.3	0.214	KIT	ML LD
	76.5	0.189	Kyoto	ML SD
	75.9	0.173	KIT	ML SD
	74.7	0.136	Kyoto	ML ZS
2	72.6	0.048	UDS-DFKI	ML SD
3	69.6	-0.070	FBK	ML SD
4	68.5	-0.103	UDS-DFKI	ML ZS
	68.1	-0.115	GTCT	ML ZS
5	60.4	-0.385	FBK	ML ZS

Source-based DA

#	Ave %	Ave z	System	Condition
1	66.1	0.165	KIT	ML SD
	65.4	0.145	Kyoto	ML ZS
	65.1	0.142	KIT	ML LD
	64.2	0.112	Kyoto	ML SD
2	61.5	0.021	UDS-DFKI	ML SD
3	60.0	-0.050	FBK	ML SD
4	58.1	-0.095	UDS-DFKI	ML ZS
	58.3	-0.102	GTCT	ML ZS
5	54.0	-0.229	FBK	ML ZS

Reference-based DA

Post-Editing

tst 2017 HE SET

10 TED Talks

- initial 50% of each talk
- 603 src sentences
- ~10K src words

Post-Editing

same dataset for
Nl-De and Ro-It

tst 2017 HE SET

10 TED Talks

- initial 50% of each talk
- 603 src sentences
- ~10K src words

Post-Editing

tst 2017 HE SET

10 TED Talks

- initial 50% of each talk
- 603 src sentences
- ~10K src words

SYS-1

SYS-2

...

SYS-*n*

Post-Editing

9 systems:
3 ML SD + 3 ML ZS + 3 BL SD

tst 2017 HE SET

10 TED Talks
- initial 50% of each talk
- 603 src sentences
- ~10K src words

SYS-1

SYS-2

...

SYS-*9*

Post-Editing

tst 2017 HE SET

10 TED Talks

- initial 50% of each talk
- 603 src sentences
- ~10K src words

SYS-1

SYS-2

SYS-3

SYS-*9*

SYS-1 Post-Edit

SYS-2 Post-Edit

SYS-3 Post-Edit

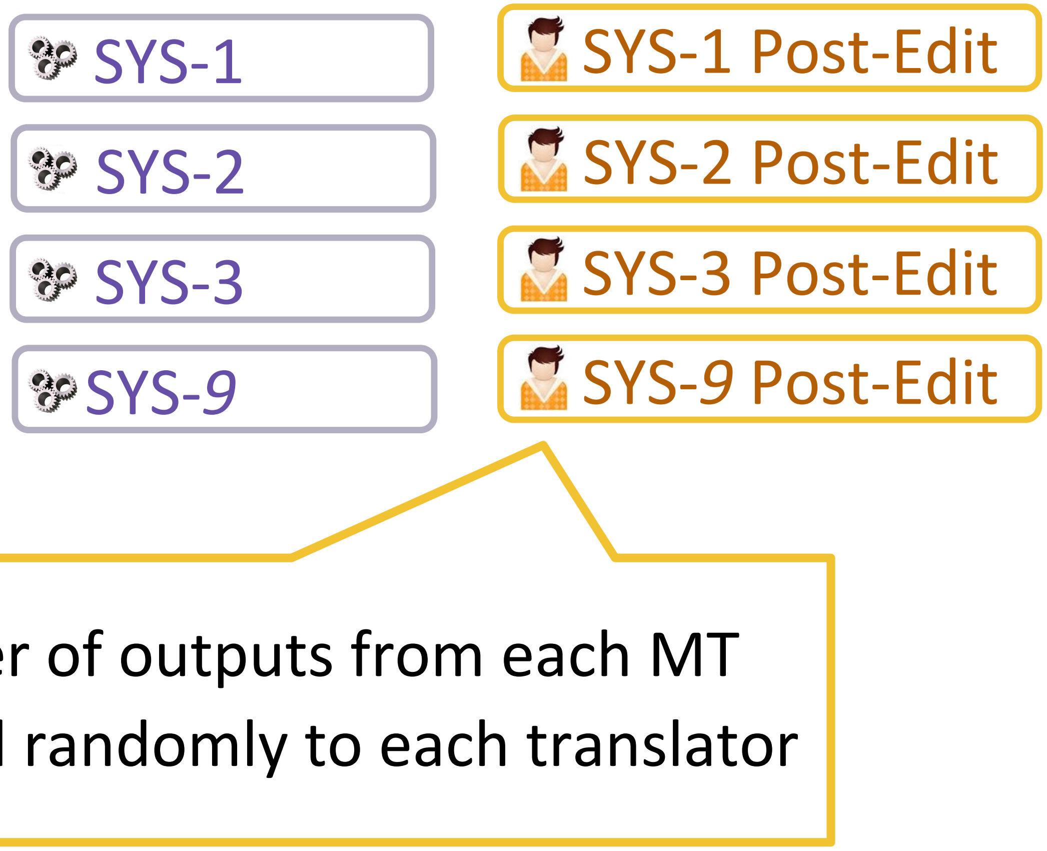
SYS-*9* Post-Edit

Post-Editing

tst 2017 HE SET

10 TED Talks

- initial 50% of each talk
- 603 src sentences
- ~10K src words



SYS-1

SYS-2

SYS-3

SYS-*9*

SYS-1 Post-Edit

SYS-2 Post-Edit

SYS-3 Post-Edit

SYS-*9* Post-Edit

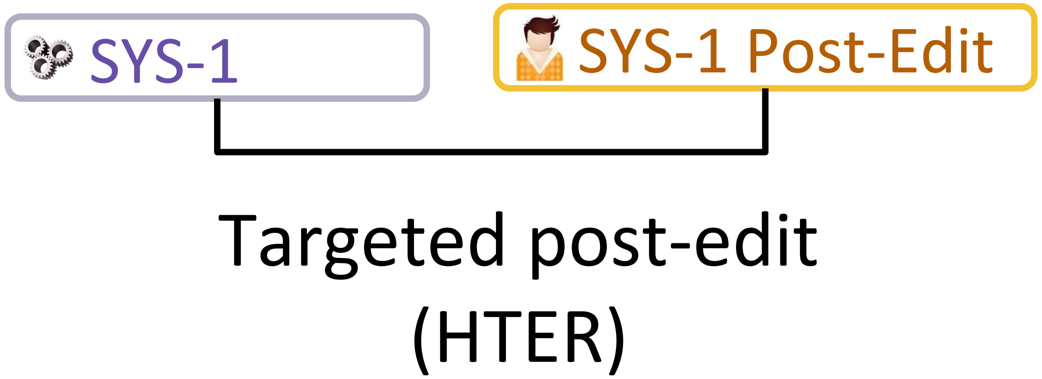
an equal number of outputs from each MT system assigned randomly to each translator

Post-Editing

tst 2017 HE SET

10 TED Talks

- initial 50% of each talk
- 603 src sentences
- ~10K src words



SYS-1

SYS-1 Post-Edit

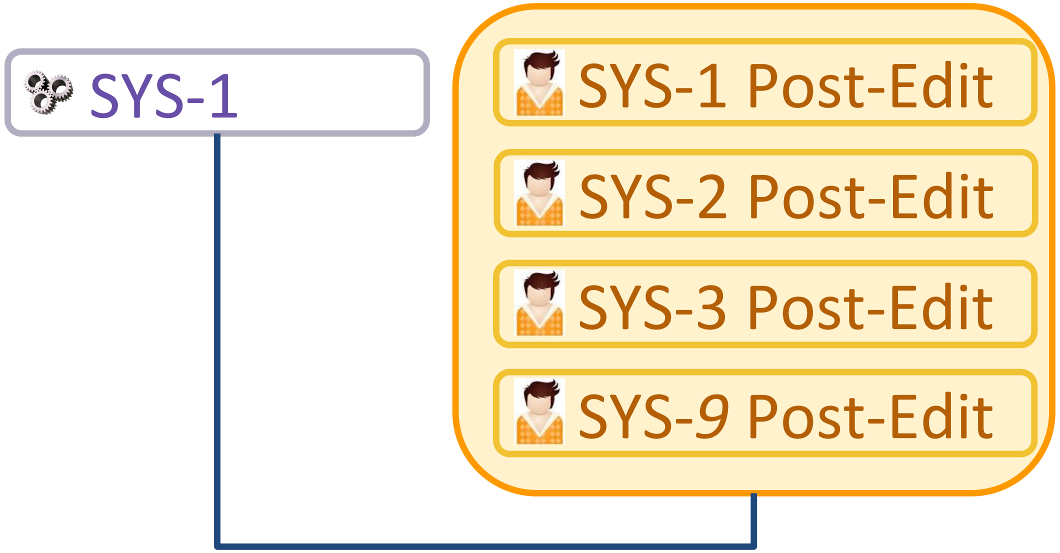
Targeted post-edit
(HTER)

Post-Editing

tst 2017 HE SET

10 TED Talks

- initial 50% of each talk
- 603 src sentences
- ~10K src words



SYS-1

SYS-1 Post-Edit

SYS-2 Post-Edit

SYS-3 Post-Edit

SYS-*9* Post-Edit

Multiple references
(mTER)

Post-Editing: Results

Nl -> De

Condition	System	mTER
ML ***ZS***	Kyoto	20.33
	FBK	26.19
	UDS-DFKI	27.36
ML ***SD***	Kyoto	20.38
	FBK	21.68
	UDS-DFKI	23.94
BL ***SD***	Kyoto	20.31
	FBK	23.71
	UDS-DFKI	30.27

Ro -> It

Condition	System	mTER
ML ***ZS***	Kyoto	22.65
	FBK	29.16
	UDS-DFKI	28.74
ML ***SD***	Kyoto	20.27
	FBK	20.74
	UDS-DFKI	23.39
BL ***SD***	Kyoto	18.39
	FBK	22.69
	UDS-DFKI	26.73

Post-Editing: Results

Nl -> De

Condition	System	mTER
ML ***ZS***	Kyoto	20.33
ML ***SD***	Kyoto	20.38
BL ***SD***	Kyoto	20.31

Post-Editing: Results

Nl -> De

Condition	System	mTER
ML ***ZS***		
ML ***SD***		
	FBK	**21.68**
	UDS-DFKI	**23.94**
BL ***SD***		
	FBK	23.71
	UDS-DFKI	30.27

Post-Editing: Results

Nl -> De

Condition	System	mTER	
ML ***ZS***			
	FBK	26.19	+4.51
	UDS-DFKI	27.36	+3.42
ML ***SD***			
	FBK	**21.68**	
	UDS-DFKI	**23.94**	
BL ***SD***			
	FBK	23.71	
	UDS-DFKI	30.27	

Post-Editing: Results

Ro -> It

Condition	System	mTER	
ML *ZS*	Kyoto	22.65	+2.38
ML *SD*	Kyoto	20.27	
BL *SD*	Kyoto	18.39	

Post-Editing: Results

Ro -> It

Condition	System	mTER
ML *ZS*		
ML *SD*		
	FBK	**20.74**
	UDS-DFKI	**23.39**
BL *SD*		
	FBK	22.69
	UDS-DFKI	26.73

Post-Editing: Results

Ro -> It

Condition	System	mTER	
ML ***ZS***			
	FBK	29.16	+8.42
	UDS-DFKI	28.74	+5.35
ML ***SD***			
	FBK	**20.74**	
	UDS-DFKI	**23.39**	
BL ***SD***			
	FBK	22.69	
	UDS-DFKI	26.73	

Final Remarks

★ Large-scale evaluation of ML/ZS translation

- ML systems are an effective alternative to BL systems
- ZS translation is feasible

Final Remarks

★ Large-scale evaluation of ML/ZS translation

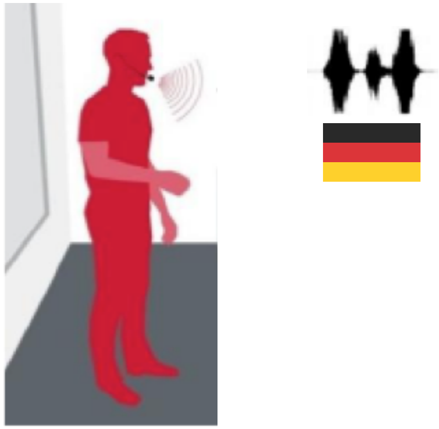
- ML systems are an effective alternative to BL systems
- ZS translation is feasible

★ Availability of a multifaceted Human Evaluation dataset

- DA: overall MT translation quality
  - *src*- vs. *ref*-based comparative analyses
- PE: MT utility in a real translation scenario
  - fine-grained analyses
  - 9 additional reference translations for each task

Lecture Task

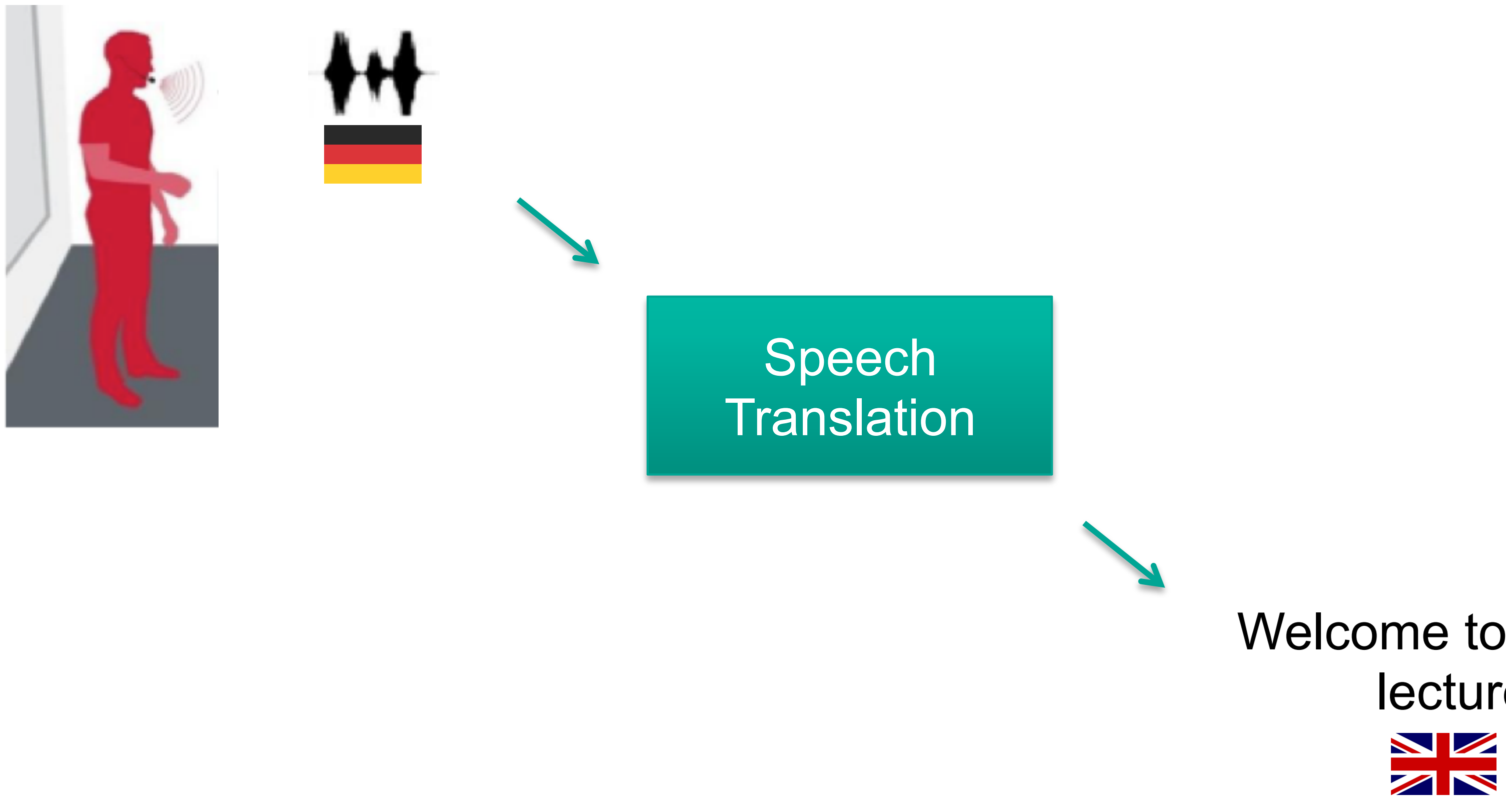
- Speech-to-Text translation task



Welcome to todays lecture

Lecture Task

- Speech-to-Text translation task

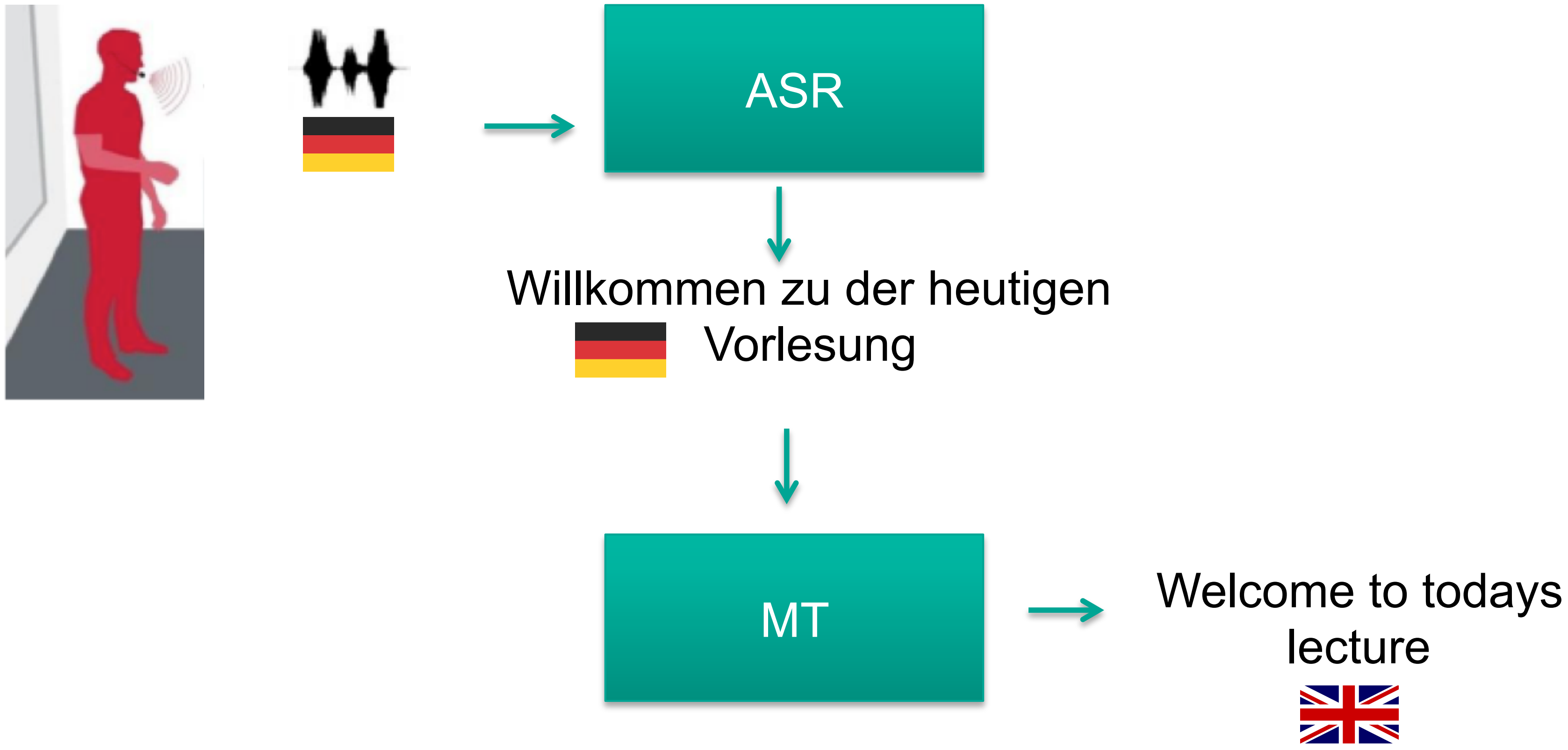


Speech Translation

Welcome to todays lecture

Lecture Task

- Speech-to-Text translation task



ASR

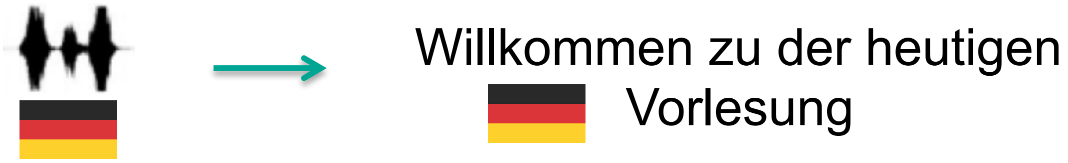
Willkommen zu der heutigen Vorlesung

MT

Welcome to todays lecture

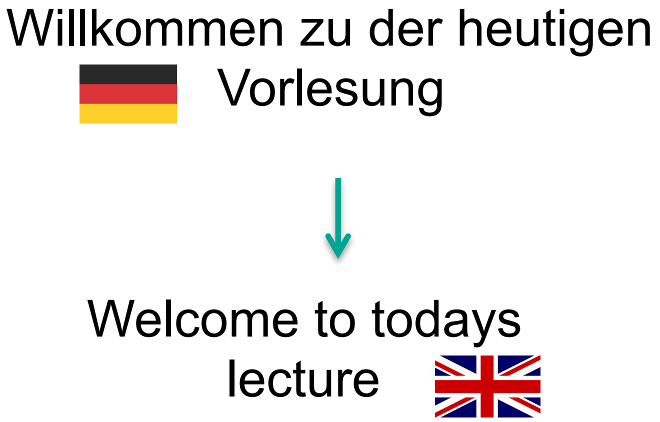
Sub-tasks

- Input:
  - Audio, not segmented
- ASR:
  - Output:
    - Text
  - Measured in WER



Willkommen zu der heutigen Vorlesung

- SLT:
  - Output:
    - Target language text
  - Measured in BLEU



Willkommen zu der heutigen Vorlesung

Welcome to todays lecture

Conditions

- German to English
  - ASR:
    - German
  - SLT:
    - German to English translation

🇩🇪 → 🇬🇧

- English to German
  - ASR:
    - German
  - SLT:
    - English to German translation

🇬🇧 → 🇩🇪

Challanges

- University lectures:
  - Specific vocabulary
  - Less prepared speech than TED talks
- Unsegmented audio
  - Segmentation for ASR
  - Segmentation for MT
    - Punctuation prediction
- Translation of speech
  - Handle noise in the ASR output

ASR Results: German

Test set	KIT
Lecture 01	16.6
Lecture 03	31.8
Lecture 04	17.7
All	21.3

ASR Results: English

Test set	KIT
Lecture 01	9.9
Lecture 02	11.7
TED 2403	6.6
TED 2429	10.6
TED 2438	6.6
TED 2439	15.5
TED 2440	4.1
TED 2442	6.7
TED 2447	6.0
TED 2507	6.2
All Lectures	10.3
All TED	7.7
All	8.5

SLT Results: German - English

Test set	KIT	UEDIN
Lecture 01	17.31	18.86
Lecture 03	7.66	8.39
Lecture 04	15.32	17.58
All	12.50	13.99

SLT Results: English - German

Test set	KIT	UEDIN
Lecture 01	23.40	23.56
Lecture 02	18.75	22.70
TED 2403	18.67	16.48
TED 2429	23.87	16.17
TED 2438	17.14	8.05
TED 2439	14.85	8.71
TED 2440	13.52	13.28
TED 2442	20.89	16.30
TED 2447	11.59	7.73
TED 2478	17.67	12.69
TED 2507	16.64	14.15
All	18.59	15.98

Dialogues Task

Katsuhito Sudoh Koichiro Yoshino

NAIST (Nara Institute of Science and Technology)
Japan

Quick Summary

- NEW task: Translating *attentive listening* dialogues
  - Japanese-to-English
  - Relatively long conversation (~300 utterances each)
  - Highly context dependent
- Only dev. and test sets were supplied
  - Participants can use any external resources for training
- NO participants in this year ☹
  - No results in this talk…

Attentive Listening

- A listener listens to people about what they think
  - Basically natural conversation
  - Many spontaneous speech phenomena (esp. disfluency)

LI: How many brothers or sisters do you have?

SP: It’s the two of us, my brother and I.

LI: A younger brother?

SP: No, I have an elder brother.

LI: Oh, really? Is he in good health?

SP: No, he has passed away already.

LI: I’m sorry to hear that…

Difficulty (Even by professinal translators…)

- Non task-oriented, open-domain
- Spontaneous speech phenomena (disfluency)
- Many context dependent utterances
- Anaphora resolution, zero pronoun

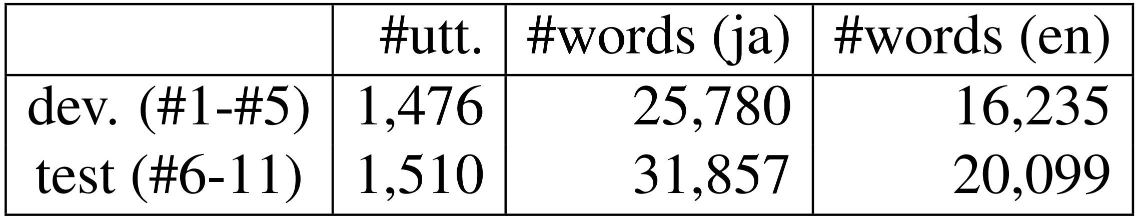
SP: No, I have an elder brother. （いや、兄です。）
LI: Oh, really? Is he in good health? （そうですか。ご健在ですか？）
SP: No, he has passed away already. （いや、もう亡くなりました。）

MT tasks in past IWSLT

- Conversation in travel situation
  - BTEC: basic experssions - for long time
  - SLDB: translator-assisted cross-lingual dialogues - 2009
  - Olympics (a.k.a. HIT corpus): short conversation – 2012
- Monologue
  - TED Talks
  - Lectures

Data (available in eval. website)

- NAIST Attentive Listening Corpus
  - H. Tanaka et al., in Proc. O-COCOSDA 2016
  - Dialogues between elderly people and listeners
  - Japanese, mostly in Kansai dialects
- Data preprocessing for dev. and test sets
  - 11 dialogues (out of 50 in the corpus)
  - Translation into English by professional translators
  - Rewriting into standard Japanese



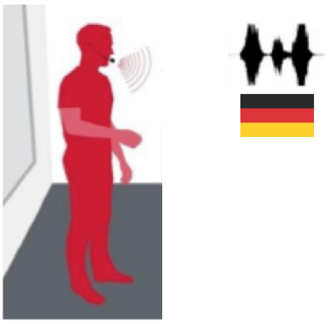
	#utt.	#words (ja)	#words (en)
dev. (#1-#5)	1,476	25,780	16,235
test (#6-11)	1,510	31,857	20,099

We’re looking forward to your challenge…!

NAIST®

Lecture Task

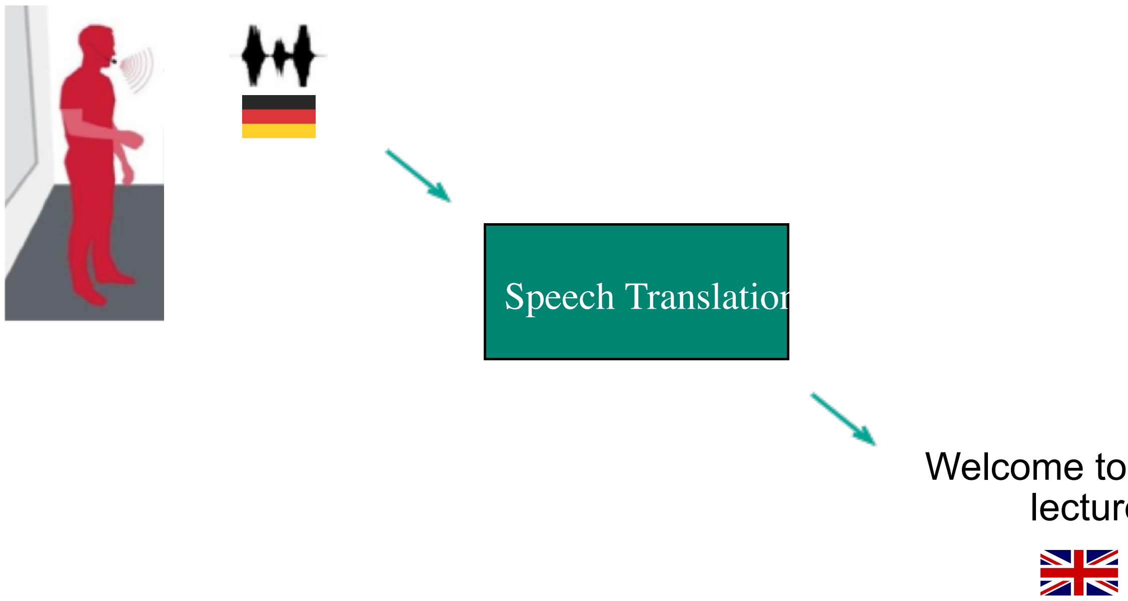
- Speech-to-Text translation task



Welcome to todays
lecture

Lecture Task

- Speech-to-Text translation task

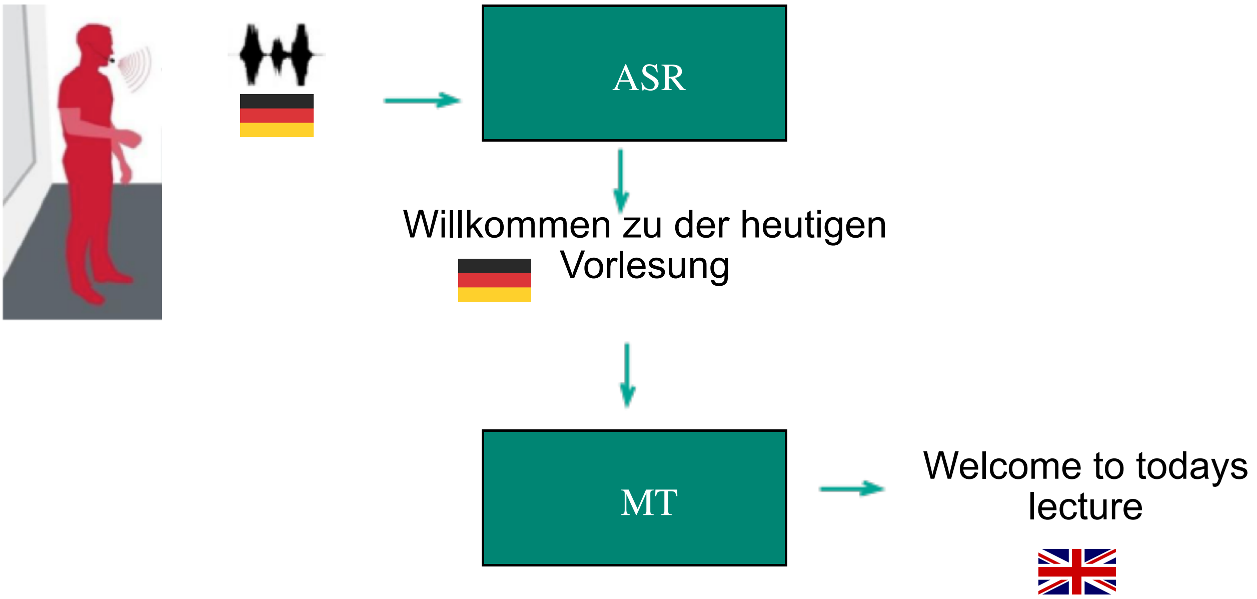


Speech Translation

Welcome to todays lecture

Lecture Task

- Speech-to-Text translation task



ASR

Willkommen zu der heutigen Vorlesung

MT

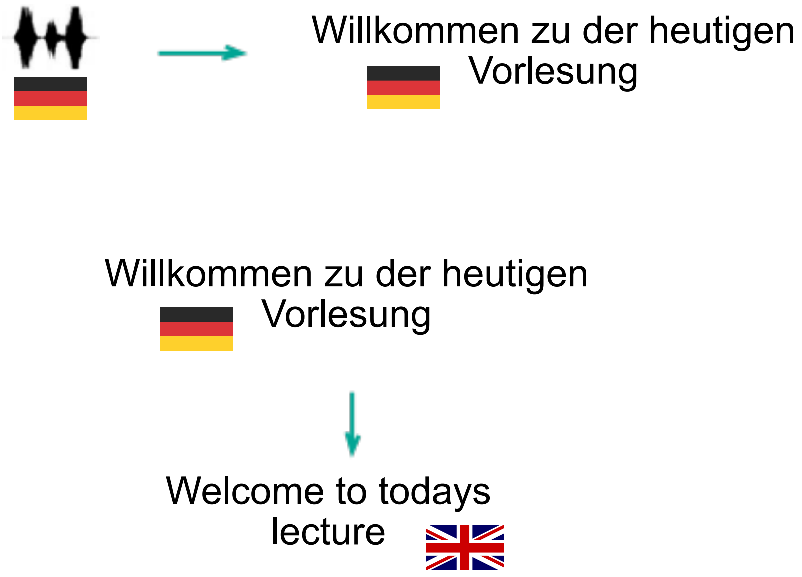
Welcome to todays lecture

Sub-tasks

- Input:
  - Audio, not segmented

- ASR:
  - Output:
    - Text
  - Measured in WER

- SLT:
  - Output:
    - Target language text
  - Measured in BLEU



Willkommen zu der heutigen Vorlesung

Willkommen zu der heutigen Vorlesung

Welcome to todays lecture

Conditions

- German to English
  - ASR:
    - German
  - SLT:
    - German to English translation

- English to German
  - ASR:
    - German
  - SLT:
    - English to German translation

Challanges

- University lectures:
  - Specific vocabulary
  - Less prepared speech than TED talks
- Unsegmented audio
  - Segmentation for ASR
  - Segmentation for MT
    - Punctuation prediction
- Translation of speech
  - Handle noise in the ASR output

ASR Results: German

Test set	KIT
Lecture 01	16.6
Lecture 03	31.8
Lecture 04	17.7
All	21.3

ASR Results: English

Test set	KIT
Lecture 01	9.9
Lecture 02	11.7
TED 2403	6.6
TED 2429	10.6
TED 2438	6.6
TED 2439	15.5
TED 2440	4.1
TED 2442	6.7
TED 2447	6.0
TED 2507	6.2
All Lectures	10.3
All TED	7.7
All	8.5

SLT Results: German - English

Test set	KIT	UEDIN
Lecture 01	17.31	18.86
Lecture 03	7.66	8.39
Lecture 04	15.32	17.58
All	12.50	13.99

SLT Results: English - German

Test set	KIT	UEDIN
Lecture 01	23.40	23.56
Lecture 02	18.75	22.70
TED 2403	18.67	16.48
TED 2429	23.87	16.17
TED 2438	17.14	8.05
TED 2439	14.85	8.71
TED 2440	13.52	13.28
TED 2442	20.89	16.30
TED 2447	11.59	7.73
TED 2478	17.67	12.69
TED 2507	16.64	14.15
All	18.59	15.98

Dialogues Task

Katsuhito Sudoh　Koichiro Yoshino

NAIST (Nara Institute of Science and Technology)
Japan

Quick Summary

- NEW task: Translating *attentive listening* dialogues
  - Japanese-to-English
  - Relatively long conversation (~300 utterances each)
  - Highly context dependent
- Only dev. and test sets were supplied
  - Participants can use any external resources for training
- NO participants in this year ☹
  - No results in this talk…

Attentive Listening

- A listener listens to people about what they think
  - Basically natural conversation
  - Many spontaneous speech phenomena (esp. disfluency)

LI: How many brothers or sisters do you have?

SP: It’s the two of us, my brother and I.

LI: A younger brother?

SP: No, I have an elder brother.

LI: Oh, really? Is he in good health?

SP: No, he has passed away already.

LI: I’m sorry to hear that…

Difficulty (Even by professinal translators...)

- Non task-oriented, open-domain
- Spontaneous speech phenomena (disfluency)
- Many context dependent utterances
- Anaphora resolution, zero pronoun

SP: No, I have an elder brother. （いや、兄です。）

LI: Oh, really? Is he in good health? （そうですか。ご健在ですか？）

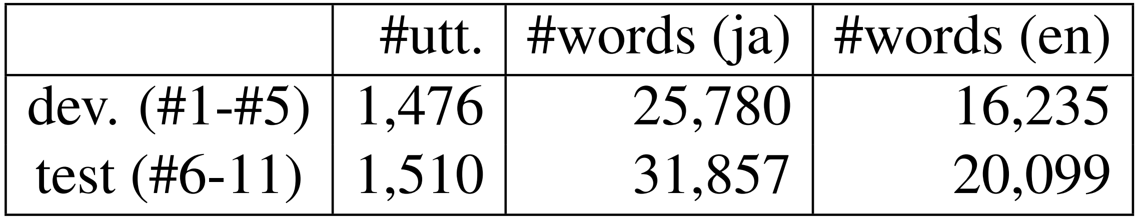
SP: No, he has passed away already. （いや、もう亡くなりました。）

MT tasks in past IWSLT

- Conversation in travel situation
  - BTEC: basic experssions - for long time
  - SLDB: translator-assisted cross-lingual dialogues - 2009
  - Olympics (a.k.a. HIT corpus): short conversation – 2012
- Monologue
  - TED Talks
  - Lectures

Data (available in eval. website)

- NAIST Attentive Listening Corpus
  - H. Tanaka et al., in Proc. O-COCOSDA 2016
  - Dialogues between elderly people and listeners
  - Japanese, mostly in Kansai dialects
- Data preprocessing for dev. and test sets
  - 11 dialogues (out of 50 in the corpus)
  - Translation into English by professional translators
  - Rewriting into standard Japanese



	#utt.	#words (ja)	#words (en)
dev. (#1-#5)	1,476	25,780	16,235
test (#6-11)	1,510	31,857	20,099

We’re looking forward to your challenge…!

NAIST®