The IWSLT 2015 Evaluation Campaign

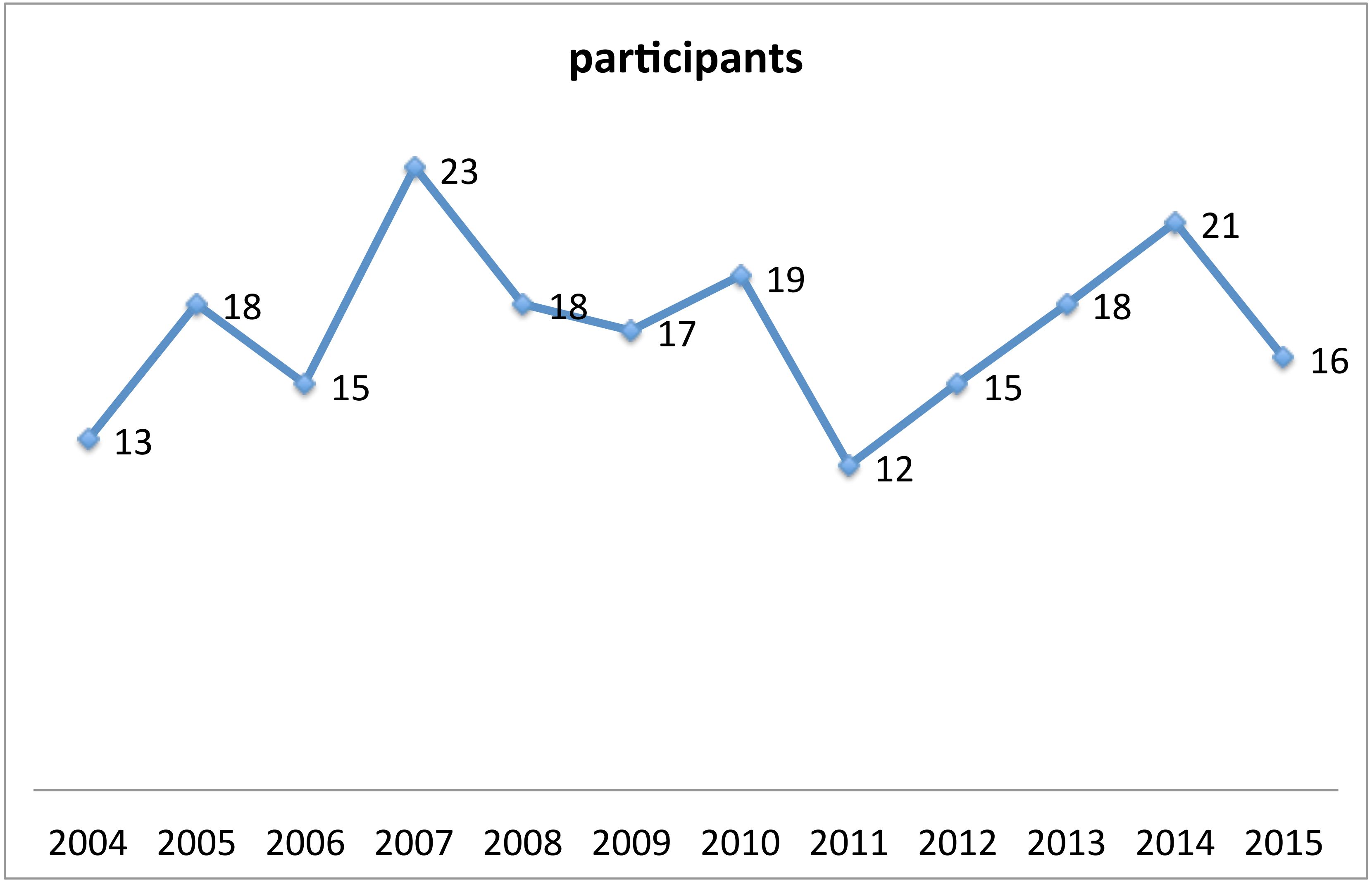
Mauro Cettolo, FBK-irst, Italy
Jan Niehues, KIT, Germany
Sebastian Stüker, KIT, Germany
Luisa Bentivogli, FBK, Italy
Roldano Cattoni, FBK, Italy
Marcello Federico, FBK-irst, Italy

IWSLT, Da Nang, 3-4 December 2015

Outline

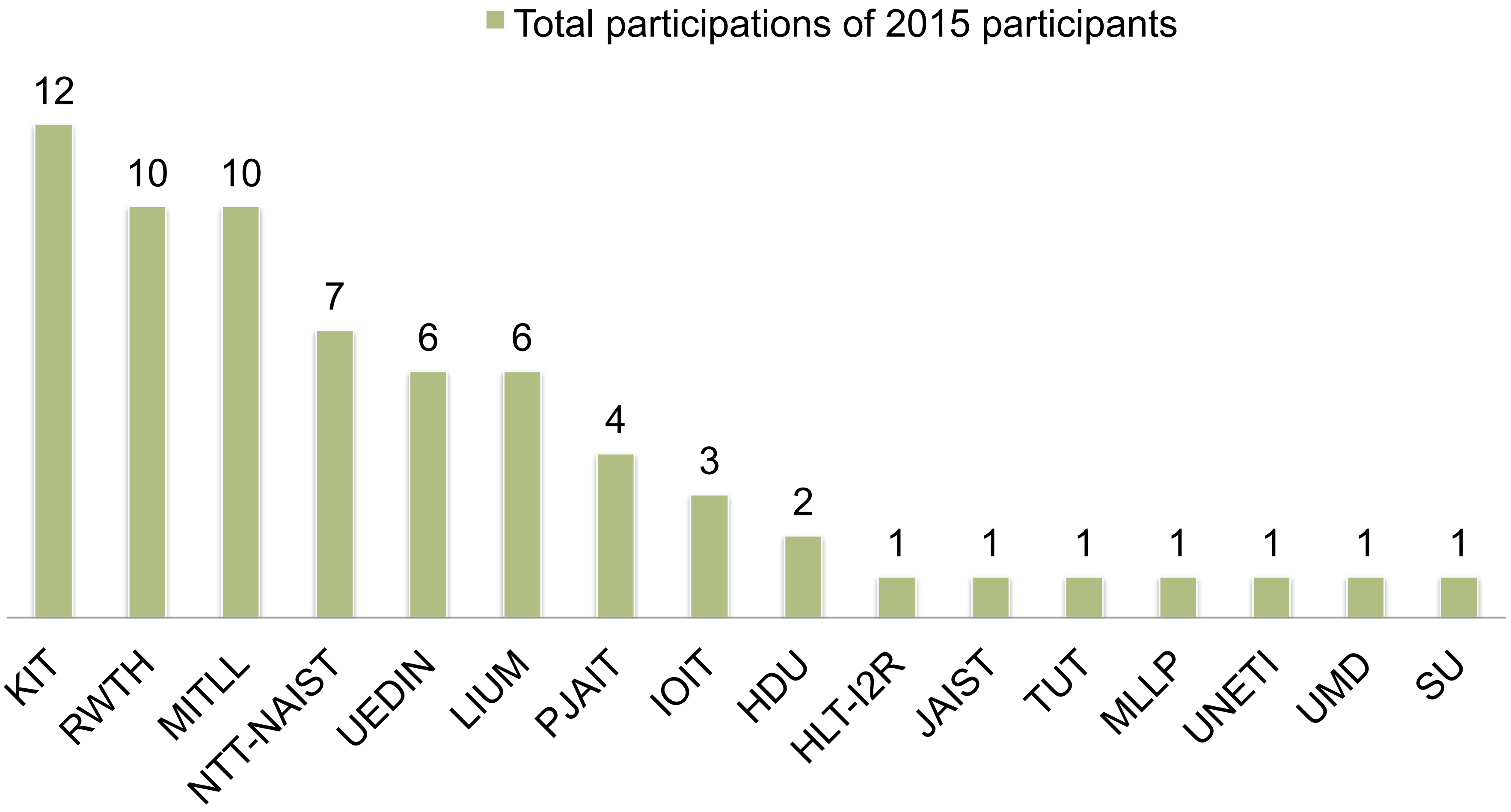
- IWSLT review
- TED Talks
- Tracks
- Automatic evaluation
- Human evaluation
- Future plans

IWSLT Evaluation: record of participants

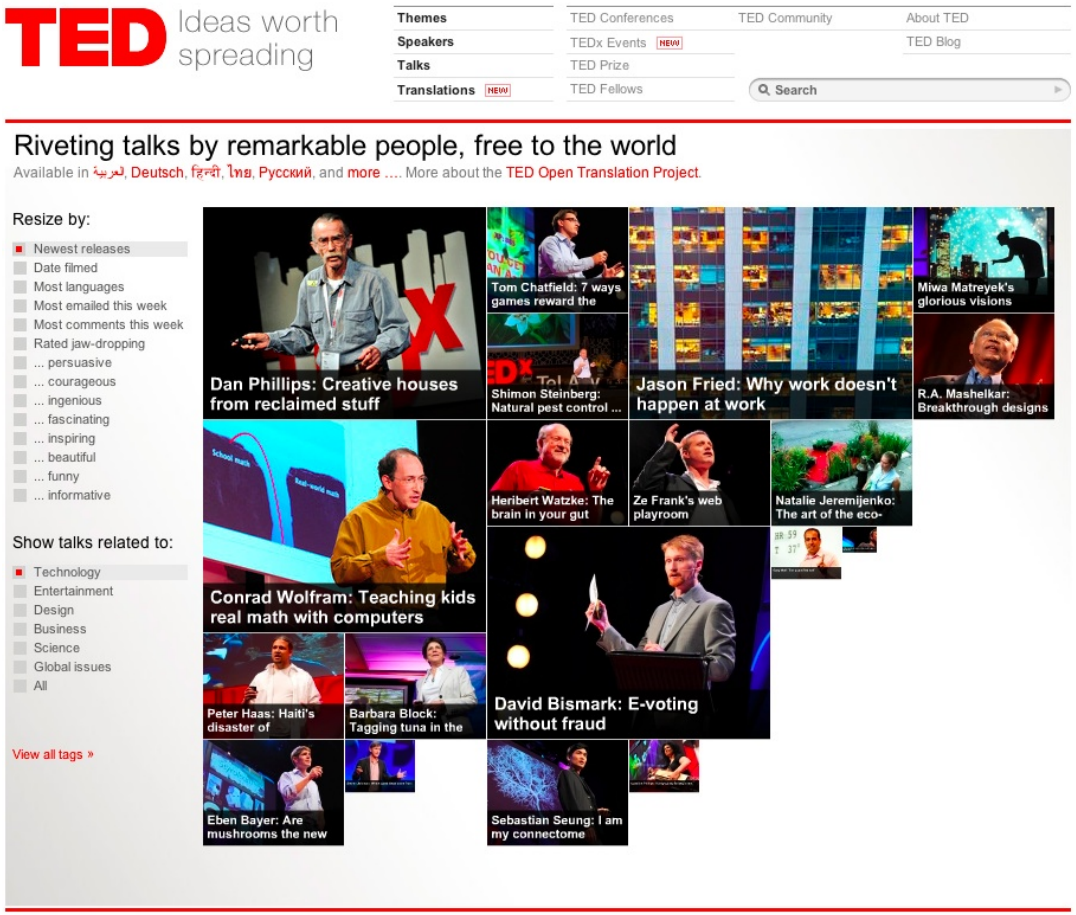


IWSLT Evaluation: record of participants

Almost 70 distinct participants in 12 years



TED Talks



TED Ideas worth spreading
Themes
Speakers
Talks
Translations NEW
TED Conferences
TEDx Events NEW
TED Prize
TED Fellows
TED Community
About TED
TED Blog
Search

Riveting talks by remarkable people, free to the world

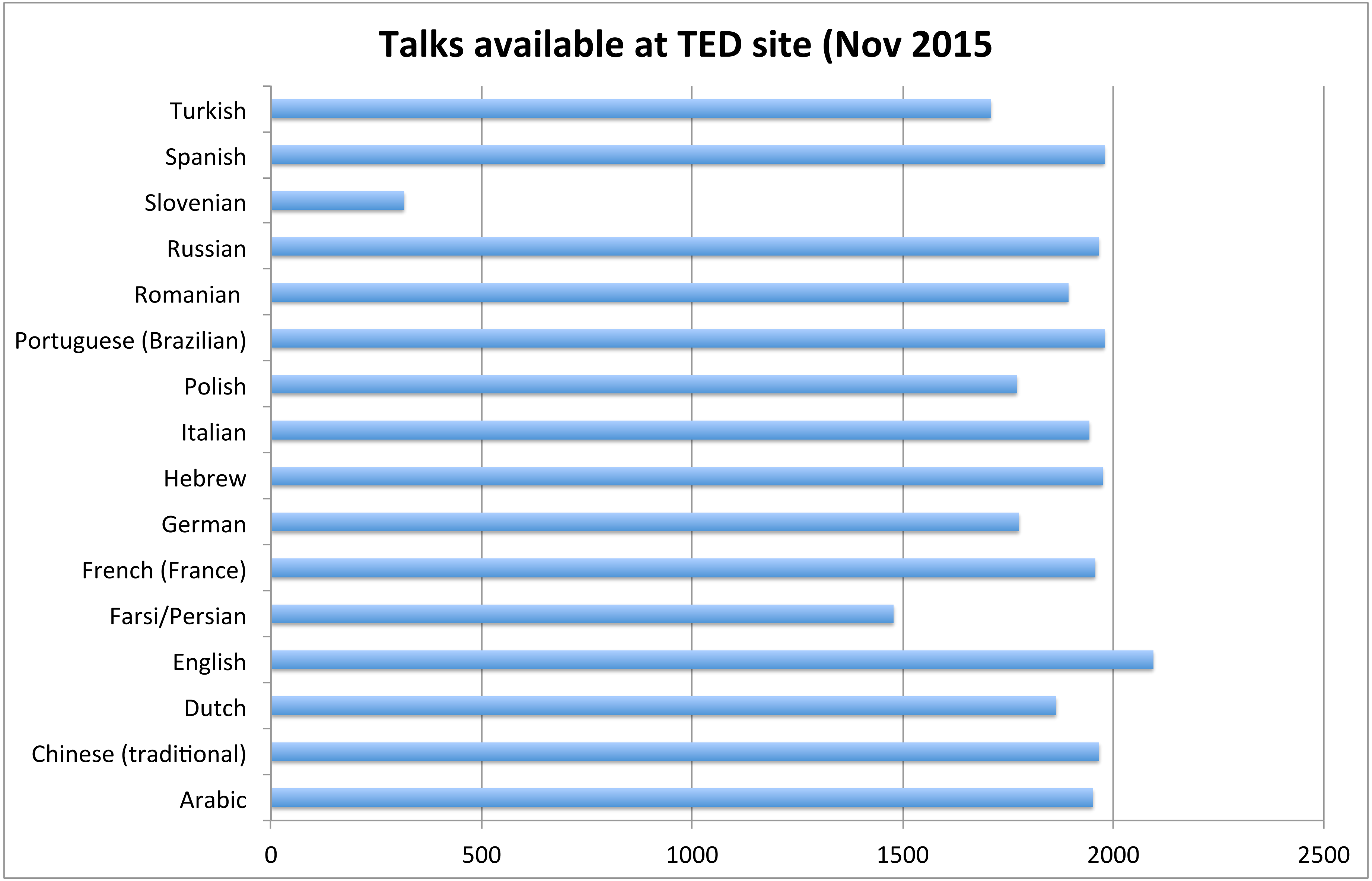
Available in عربية, Deutsch, हिन्दी, ไทย, Русский, and more .... More about the TED Open Translation Project.

Resize by:
Newest releases
Date filmed
Most languages
Most emailed this week
Most comments this week
Rated jaw-dropping
... persuasive
... courageous
... ingenious
... fascinating
... inspiring
... beautiful
... funny
... informative
Show talks related to:
Technology
Entertainment
Design
Business
Science
Global issues
All
View all tags »
Dan Phillips: Creative houses from reclaimed stuff
Tom Chatfield: 7 ways games reward the
Shimon Steinberg: Natural pest control ...
Jason Fried: Why work doesn't happen at work
Miwa Matreyek's glorious visions
R.A. Mashelkar: Breakthrough designs
Conrad Wolfram: Teaching kids real math with computers
Heribert Watzke: The brain in your gut
Ze Frank's web playroom
Natalie Jeremijenko: The art of the eco-
Peter Haas: Haiti's disaster of
Barbara Block: Tagging tuna in the
David Bismark: E-voting without fraud
Eben Bayer: Are mushrooms the new
Sebastian Seung: I am my connectome

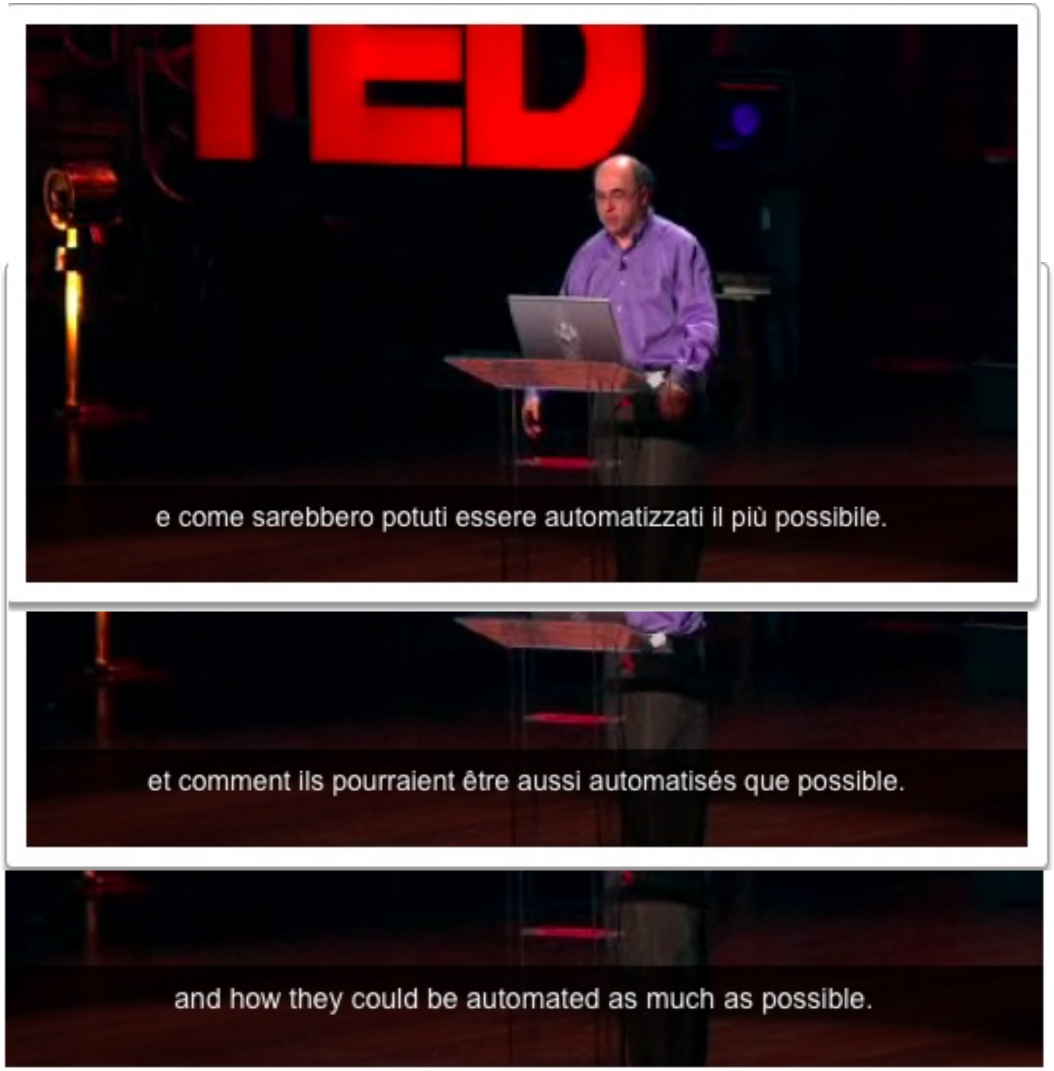
- TED LLC is non-profit
- Two annual events
- Short talks
- Variety of topics
- Website with:
  - Videos
  - Transcripts
  - Translations
- CC License

TED Talks Translations

	Nov ‘10	Nov ‘11	Nov ‘12	Nov ‘13	Nov ‘14	Nov ‘15
Talks (EN)	800	1,080	1,395	~1,650	1,875	2,095
Languages	80	83	93	103	105	109
Translators	4,000	6,823	8,382	11,010	18,699	15,487
Translations	12,500	24,287 +94%	32,707 +34%	49,607 +52%	65,290 +32%	83,265 +28%


Talks available at TED site (Nov 2015
Turkish
Spanish
Slovenian
Russian
Romanian
Portuguese (Brazilian)
Polish
Italian
Hebrew
German
French (France)
Farsi/Persian
English
Dutch
Chinese (traditional)
Arabic
0
500
1000
1500
2000
2500

Human task: subtitling and translating



e come sarebbero potuti essere automatizzati il più possibile.

et comment ils pourraient être aussi automatisés que possible.

and how they could be automated as much as possible.

- ✓ segment audio
- ✓ transcribe and annotate
- ✓ split into captions
- ✓ translate captions

Challenges in TED Task

- **Language modelling**
  - Limited in-domain training data
  - Variability of topics and styles
- **Acoustic modelling**
  - Speaker: accent, fluency, speaking rate, style, , ...
  - Noise: mumble, applauses, laughs, music, ...
- **Translation modelling**
  - Distant and under-resourced languages
  - Morphologically rich languages
- **Speech Translation**
  - From spontaneous speech to polished text
  - Detection and removal of non-speech events
  - Subtitling and translating in real-time

Challenges for 2011

- **Language modelling**
  - Limited in-domain training data
  - Variability of topics and styles
- **Acoustic modelling**
  - Speaker: accent, fluency, speaking rate, style, , ...
  - Noise: mumble, applauses, laughs, music, ...
- **Translation modelling**
  - Distant and under-resourced languages
  - Morphologically rich languages
- **Speech Translation**
  - From spontaneous speech to polished text
  - Detection and removal of non-speech events
  - Subtitling and translating a data stream in real-time

Challenges for 2012

- **Language modelling**
  - Limited in-domain training data
  - Variability of topics and styles
- **Acoustic modelling**
  - Speaker: accent, fluency, speaking rate, style, , ...
  - Noise: mumble, applauses, laughs, music, ...
- **Translation modelling**
  - Distant and **under-resourced** languages
  - **Morphologically rich languages**
- **Speech Translation**
  - From spontaneous speech to polished text
  - Detection and removal of non-speech events
  - Subtitling and translating a data stream in real-time

Challenges for 2013-2014

- **Language modelling**
  - Limited in-domain training data
  - Variability of topics and styles
- **Acoustic modelling**
  - Speaker: accent, fluency, speaking rate, style, , ...
  - **Noise: mumble, applauses, laughs, music, ...**
- **Translation modelling**
  - Distant and under-resourced languages
  - Morphologically rich languages
- **Speech Translation**
  - From spontaneous speech to polished text
  - **Detection and removal of non-speech events**
  - Subtitling and translating a data stream in real-time

Challenges for 2014-2015

- **Language modelling**
  - Limited in-domain training data
  - Variability of topics and styles
- **Acoustic modelling**
  - Speaker: accent, fluency, speaking rate, style, , ...
  - Noise: mumble, applauses, laughs, music, ...
- **Translation modelling**
  - Distant and under-resourced languages
  - Morphologically rich languages
- **Speech Translation**
  - From spontaneous speech to polished text
  - Detection and removal of non-speech events
  - **Subtitling and translating a data stream** in real-time

2015 Tracks

- **Automatic Speech Recognition (ASR)**
  - Transcription of talks from audio to text
  - English (TED), German (TEDx)
- **Spoken Language Translation (SLT)**
  - Translation of talks from audio (or ASR output) to text
  - German → English (TEDx)
  - English → Chinese, Czech, French, German, Thai, Vietnamese (TED)
- **Machine Translation (MT)**
  - Translation of talks from text to text
  - German → English (TEDx)
  - English ↔ Chinese, Czech, French, German, Thai, Vietnamese (TED)

Specifications

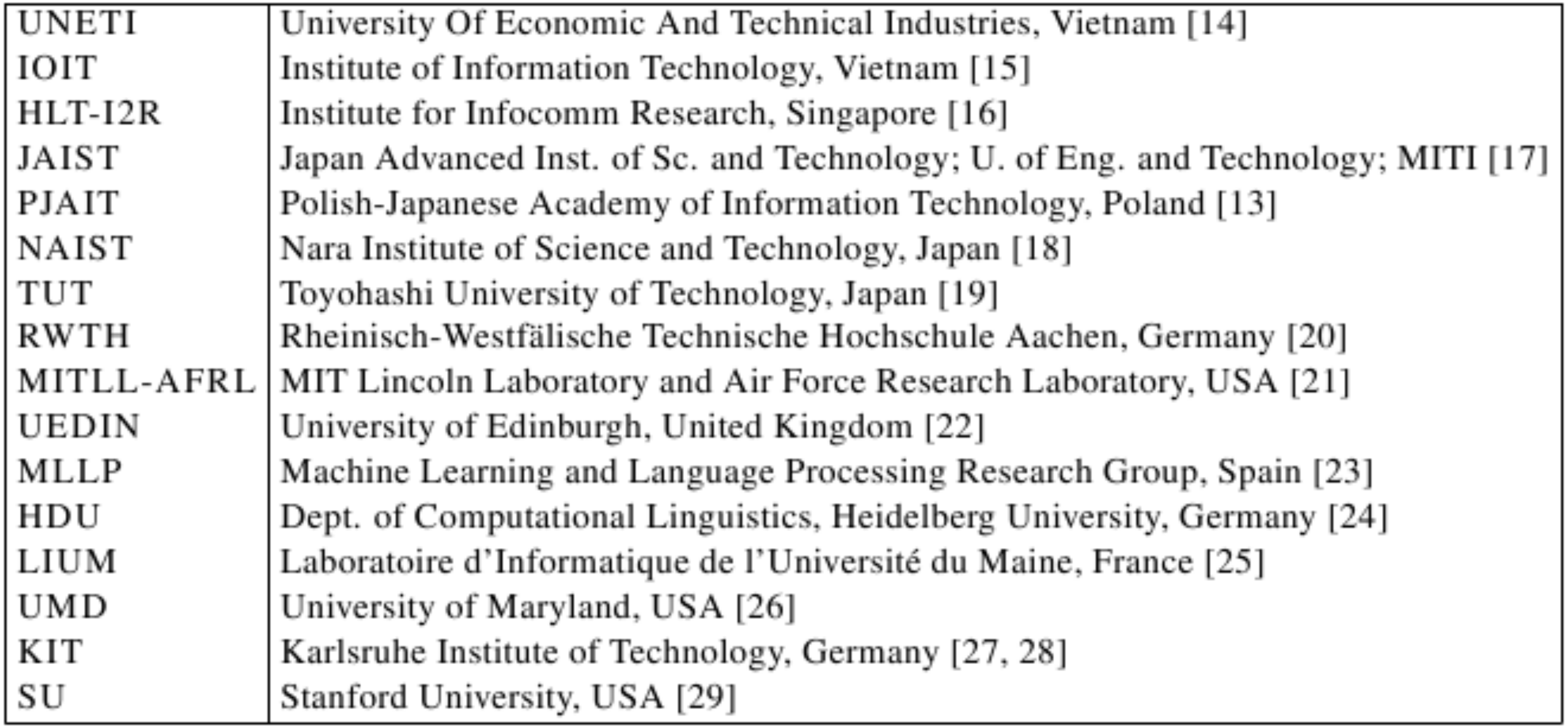
Conditions	ASR	SLT	MT
Input: Pre-segmented	no	**no**	yes
Input: Cased & Punctuated		no	yes
Output: Cased & Punctuated	no	yes	yes
Automatic evaluation	yes	yes	yes[1]
Human eval (En-Fr/De)			yes

NEW

Metrics	ASR	SLT	MT
WER	✔	✔	✔
BLEU		✔	✔
TER		✔	✔

[1] Non trivial reference baselines prepared for all directions.

Participants

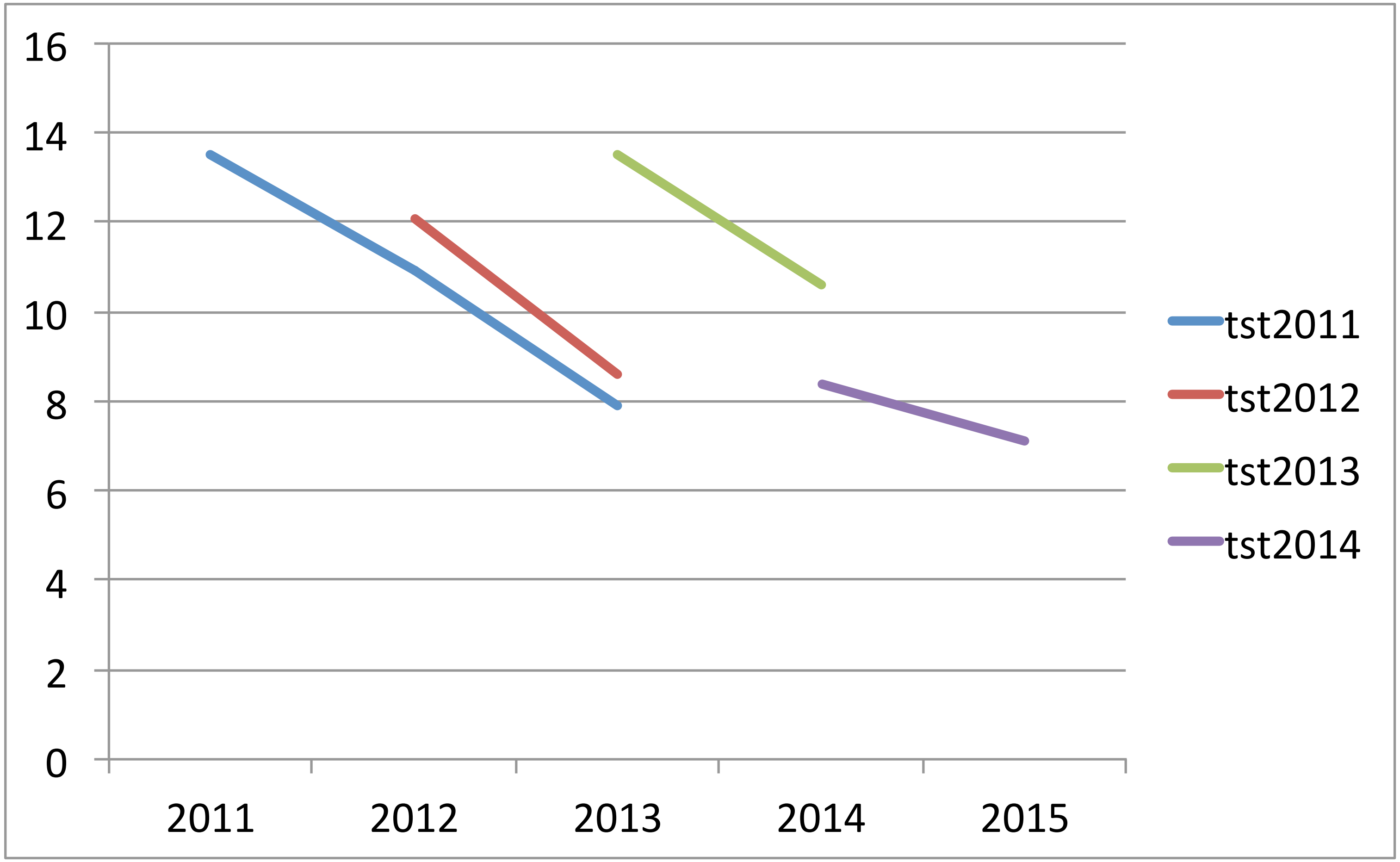


UNETI	University Of Economic And Technical Industries, Vietnam [14]
IOIT	Institute of Information Technology, Vietnam [15]
HLT-I2R	Institute for Infocomm Research, Singapore [16]
JAIST	Japan Advanced Inst. of Sc. and Technology; U. of Eng. and Technology; MITI [17]
PJAIT	Polish-Japanese Academy of Information Technology, Poland [13]
NAIST	Nara Institute of Science and Technology, Japan [18]
TUT	Toyohashi University of Technology, Japan [19]
RWTH	Rheinisch-Westfälische Technische Hochschule Aachen, Germany [20]
MITLL-AFRL	MIT Lincoln Laboratory and Air Force Research Laboratory, USA [21]
UEDIN	University of Edinburgh, United Kingdom [22]
MLLP	Machine Learning and Language Processing Research Group, Spain [23]
HDU	Dept. of Computational Linguistics, Heidelberg University, Germany [24]
LIUM	Laboratoire d'Informatique de l'Université du Maine, France [25]
UMD	University of Maryland, USA [26]
KIT	Karlsruhe Institute of Technology, Germany [27, 28]
SU	Stanford University, USA [29]

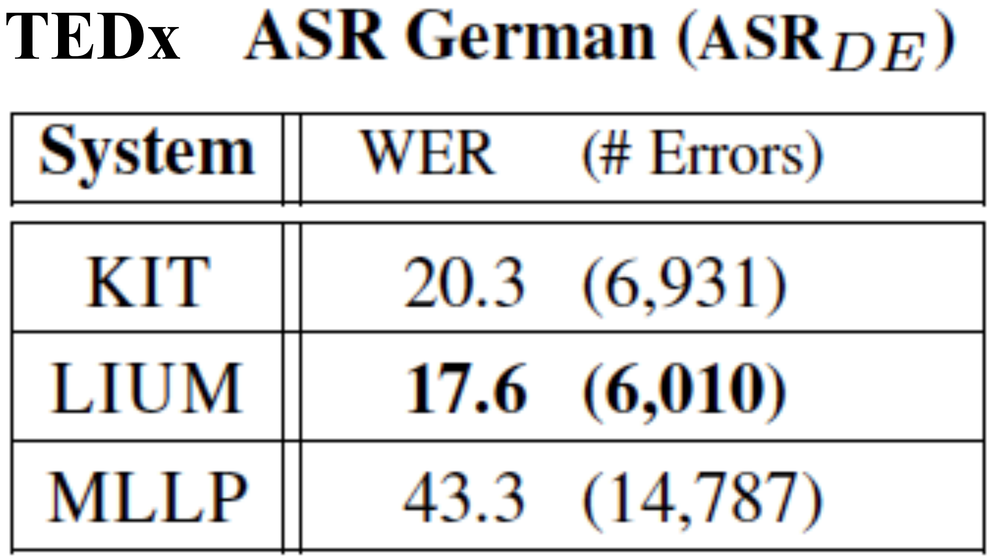
Results: ASR English (WER%)

	IWSLT15		IWSLT14		IWSLT13
	tst2015	tst2014	tst2014	tst2013	tst2013
MITLL-AFR	6.6	7.1	9.9	13.7	15.9
HLT-I2R	7.7	8.9	-	-	-
KIT	9.2	9.7	11.4	14.2	14.4
NAIST	12.0	10.4	-	-	-
MLLP	13.3	19.5	-	-	-
IOIT	13.8	13.9	19.7	24.0	27.2

Progress in ASR En (best systems WER%)



Results: ASR German

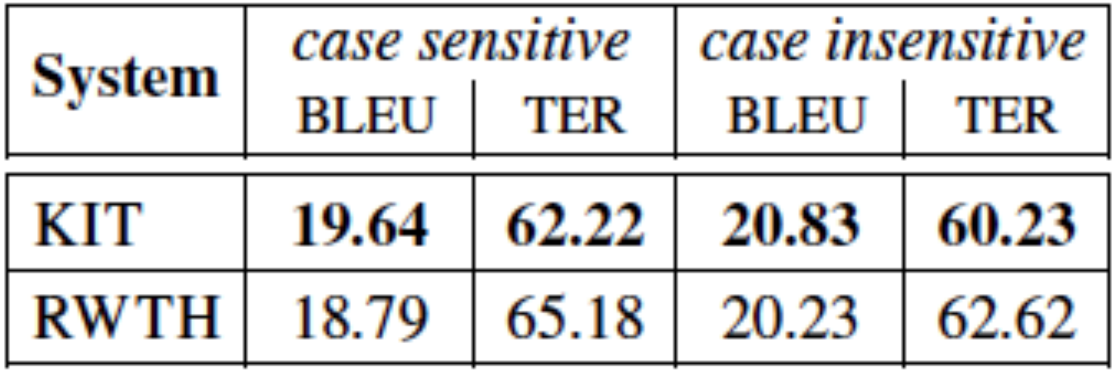


TEDx **ASR German (ASR$_{DE}$)**

System	WER (# Errors)
KIT	20.3 (6,931)
LIUM	**17.6 (6,010)**
MLLP	43.3 (14,787)

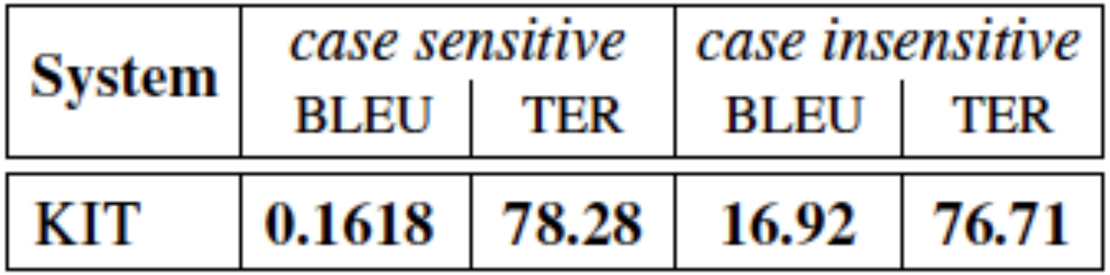
Results: SLT

TEDx : SLT German-English (MT$_{DeEn}$)



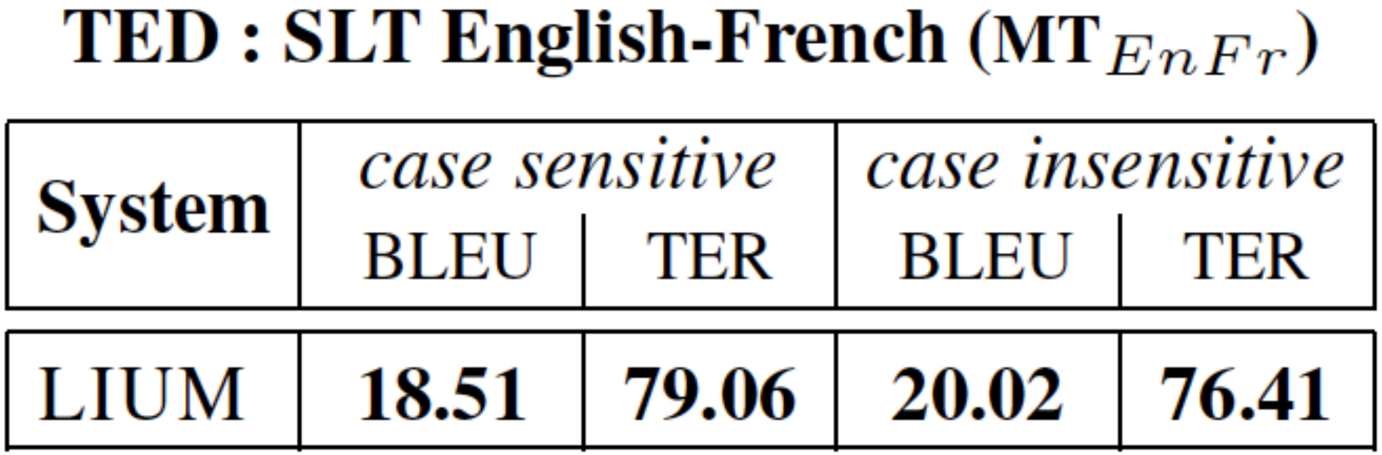
System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	**19.64**	**62.22**	**20.83**	**60.23**
RWTH	18.79	65.18	20.23	62.62

TED : SLT English-German (MT$_{EnDe}$)



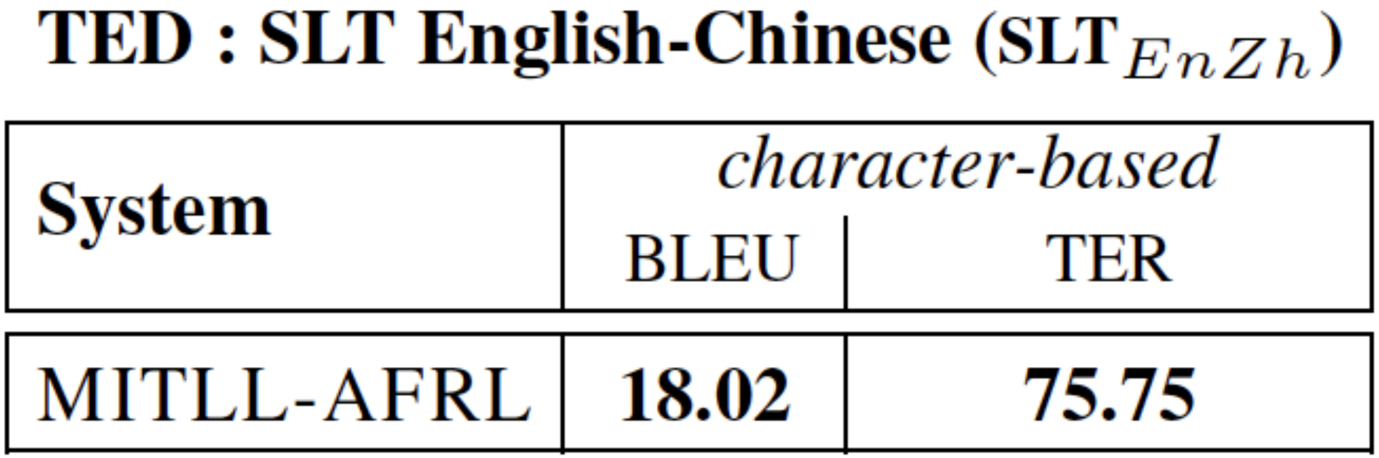
System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	**0.1618**	**78.28**	**16.92**	**76.71**

Results: SLT



TED : SLT English-French (MT$_{EnFr}$)

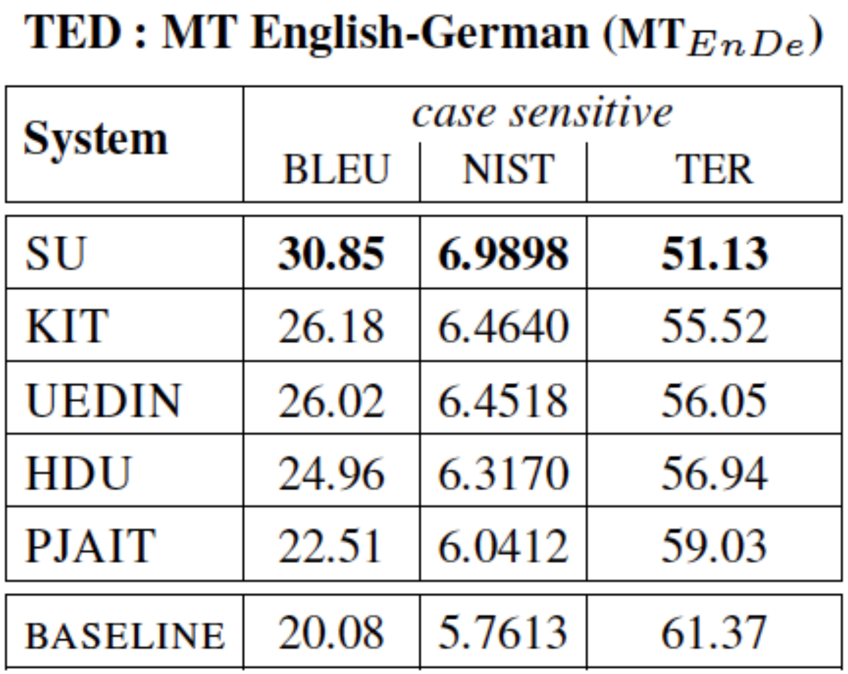
System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
LIUM	**18.51**	**79.06**	**20.02**	**76.41**



TED : SLT English-Chinese (SLT$_{EnZh}$)

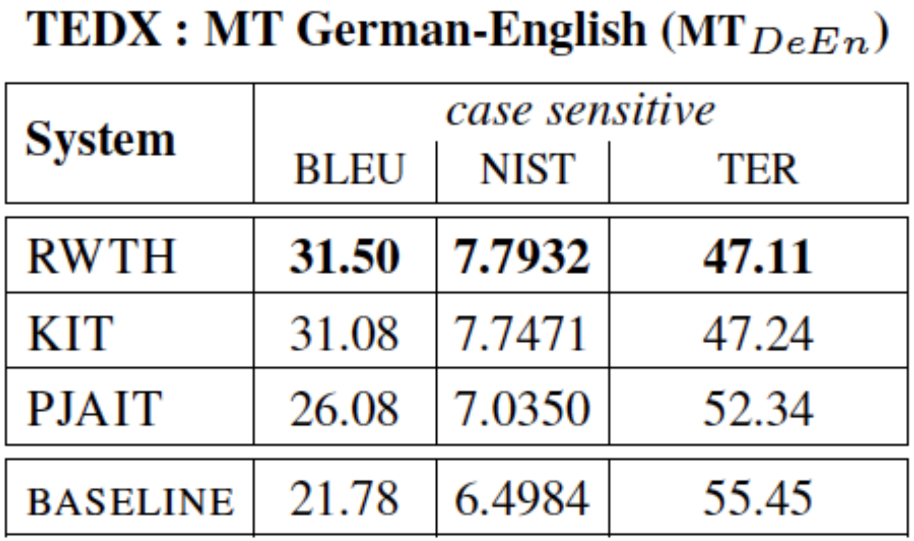
System	*character-based*	
	BLEU	TER
MITLL-AFRL	**18.02**	**75.75**

Results: MT



TED : MT English-German (MT$_{EnDe}$)

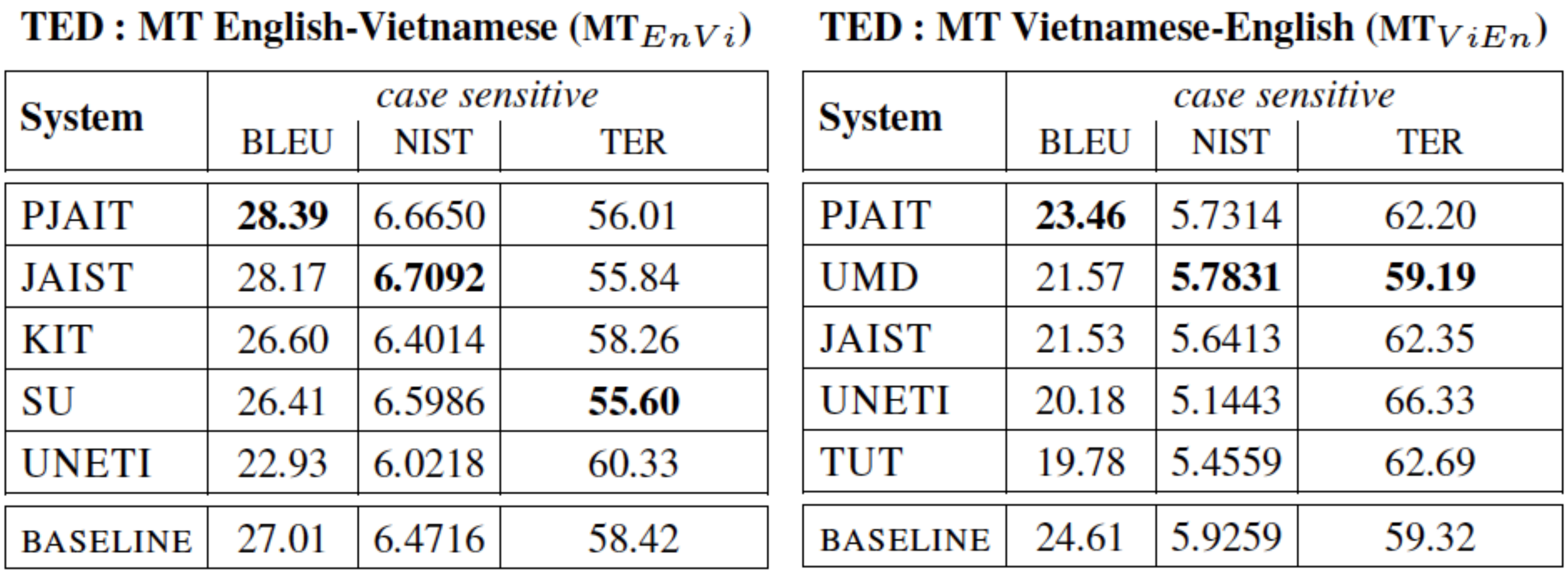
System	*case sensitive*		
	BLEU	NIST	TER
SU	**30.85**	**6.9898**	**51.13**
KIT	26.18	6.4640	55.52
UEDIN	26.02	6.4518	56.05
HDU	24.96	6.3170	56.94
PJAIT	22.51	6.0412	59.03
BASELINE	20.08	5.7613	61.37



TEDX : MT German-English (MT$_{DeEn}$)

System	*case sensitive*		
	BLEU	NIST	TER
RWTH	**31.50**	**7.7932**	**47.11**
KIT	31.08	7.7471	47.24
PJAIT	26.08	7.0350	52.34
BASELINE	21.78	6.4984	55.45

Results: MT



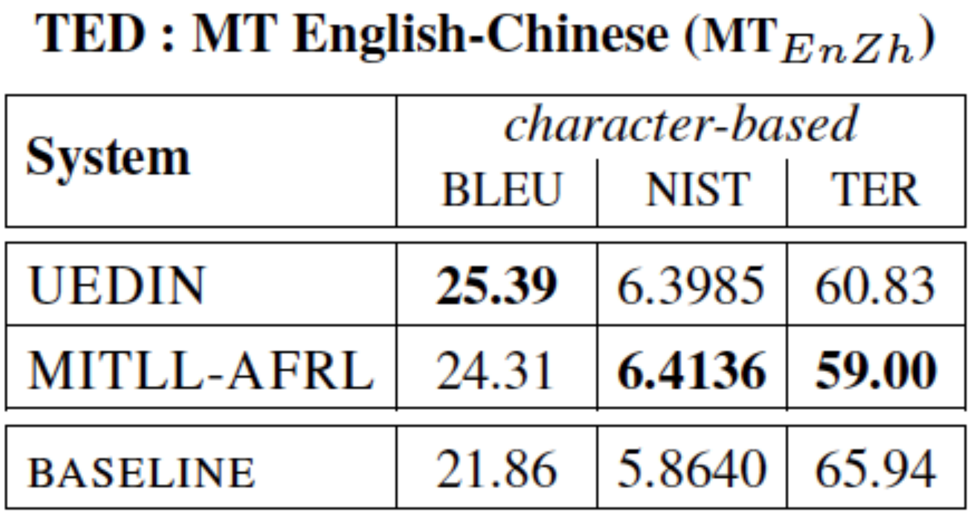
TED : MT English-Vietnamese (MT$_{EnVi}$)

System	*case sensitive* BLEU	NIST	TER
PJAIT	**28.39**	6.6650	56.01
JAIST	28.17	**6.7092**	55.84
KIT	26.60	6.4014	58.26
SU	26.41	6.5986	**55.60**
UNETI	22.93	6.0218	60.33
BASELINE	27.01	6.4716	58.42

TED : MT Vietnamese-English (MT$_{ViEn}$)

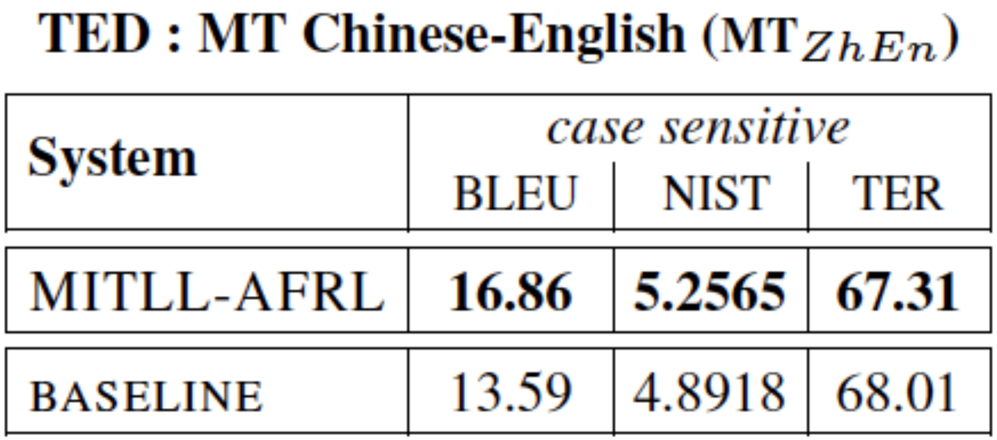
System	*case sensitive* BLEU	NIST	TER
PJAIT	**23.46**	5.7314	62.20
UMD	21.57	**5.7831**	**59.19**
JAIST	21.53	5.6413	62.35
UNETI	20.18	5.1443	66.33
TUT	19.78	5.4559	62.69
BASELINE	24.61	5.9259	59.32

Results: MT



TED : MT English-Chinese (MT_{EnZh})

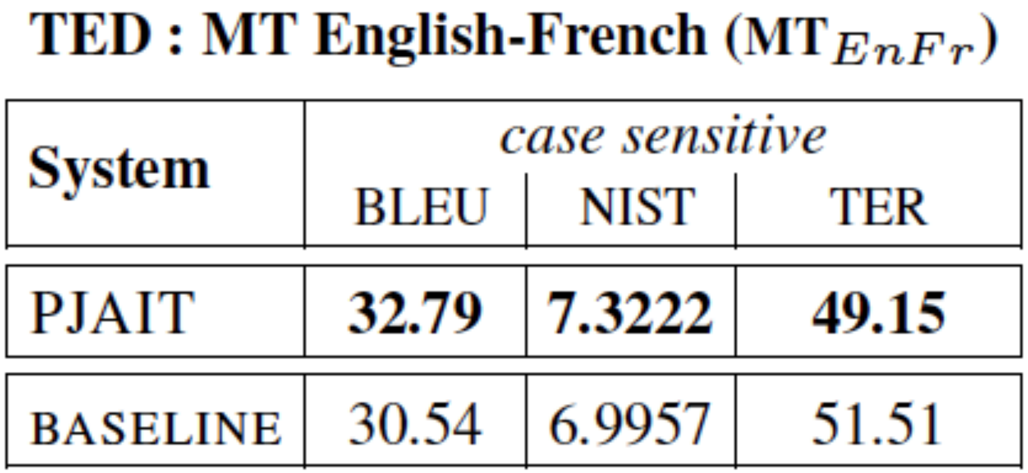
System	*character-based*		
	BLEU	NIST	TER
UEDIN	**25.39**	6.3985	60.83
MITLL-AFRL	24.31	**6.4136**	**59.00**
BASELINE	21.86	5.8640	65.94



TED : MT Chinese-English (MT_{ZhEn})

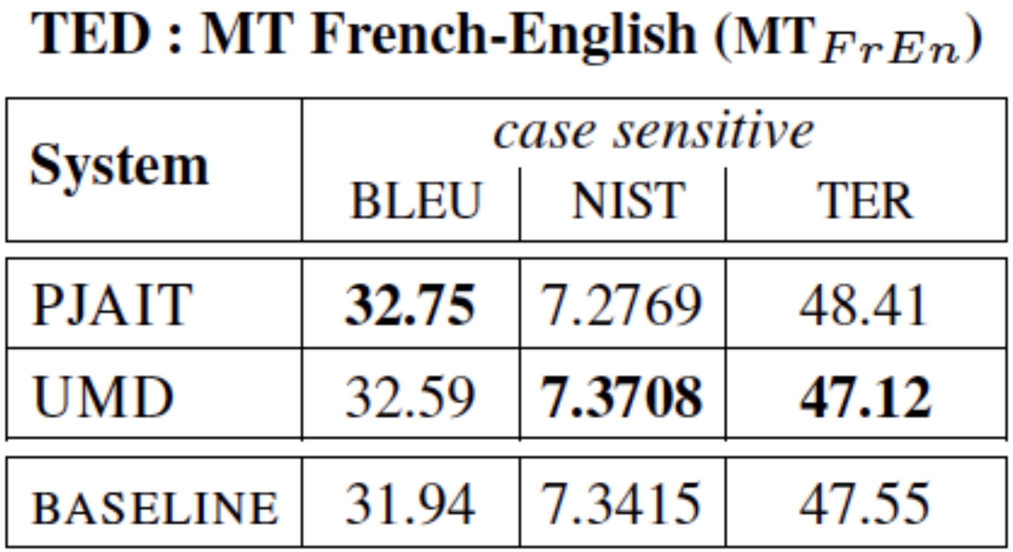
System	*case sensitive*		
	BLEU	NIST	TER
MITLL-AFRL	**16.86**	**5.2565**	**67.31**
BASELINE	13.59	4.8918	68.01

Results: MT



TED : MT English-French (MT$_{EnFr}$)

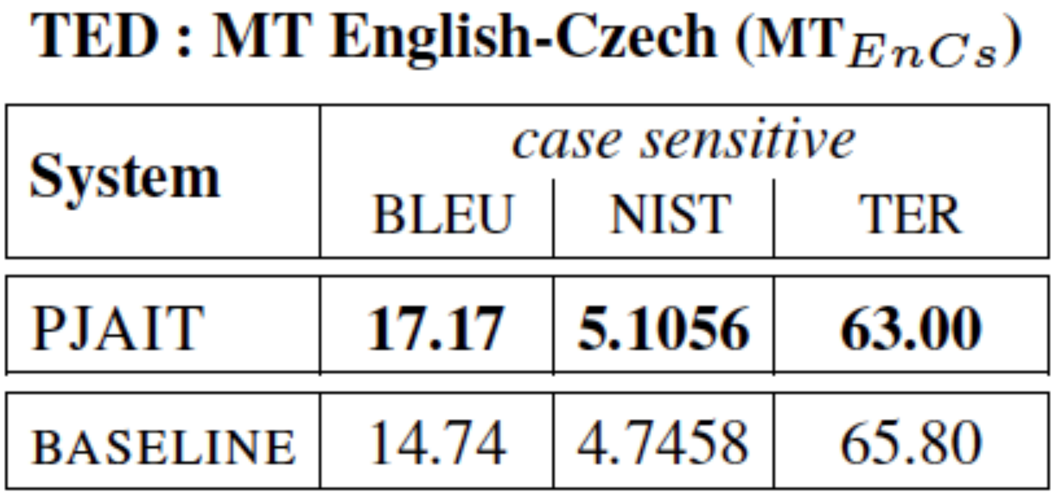
System	*case sensitive*		
	BLEU	NIST	TER
PJAIT	**32.79**	**7.3222**	**49.15**
BASELINE	30.54	6.9957	51.51



TED : MT French-English (MT$_{FrEn}$)

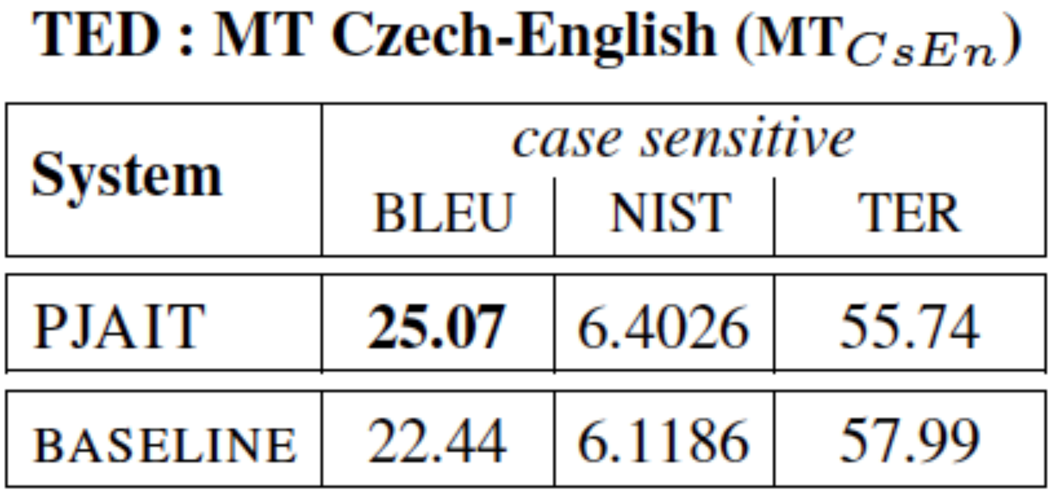
System	*case sensitive*		
	BLEU	NIST	TER
PJAIT	**32.75**	7.2769	48.41
UMD	32.59	**7.3708**	**47.12**
BASELINE	31.94	7.3415	47.55

Results: MT



TED : MT English-Czech (MT_{EnCs})

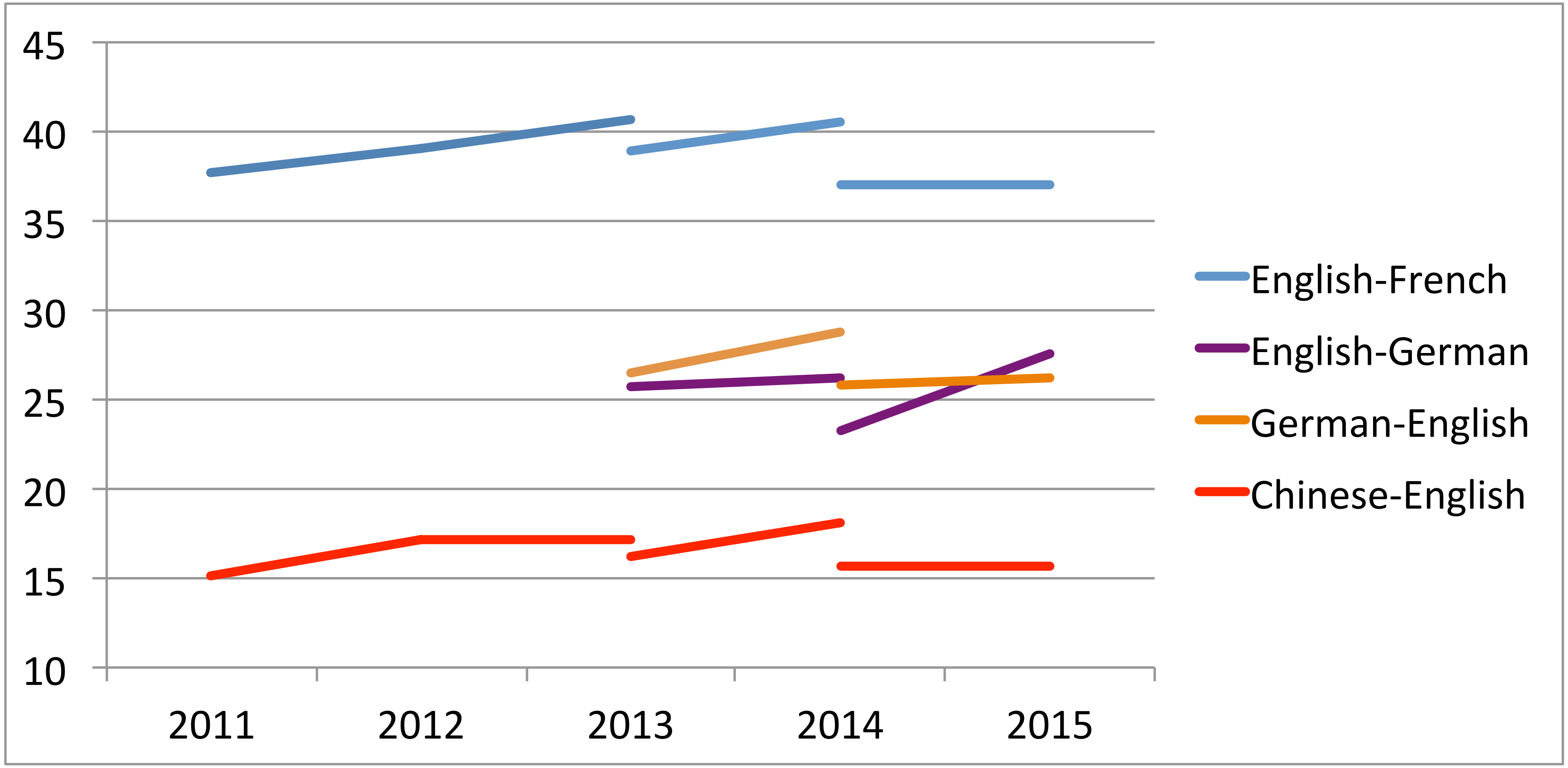
System	*case sensitive*		
	BLEU	NIST	TER
PJAIT	**17.17**	**5.1056**	**63.00**
BASELINE	14.74	4.7458	65.80



TED : MT Czech-English (MT_{CsEn})

System	*case sensitive*		
	BLEU	NIST	TER
PJAIT	**25.07**	6.4026	55.74
BASELINE	22.44	6.1186	57.99

Progress in MT (best systems BLEU%)



Human Evaluation

- Following IWSLT 2013/14: ***Post-Editing + HTER***
  - TED task as an interesting application scenario to test the utility of MT systems in a real subtitling task
  - Additional reference translations
  - Edits point to specific translation errors
  - HTER correlates well with human judgments

- Evaluation of **MT-*EnDe*** and **MT-*ViEn*** tasks

- Performed on 2015 test set (*tst2015*)

Evaluation Dataset

Human Evaluation (HE) Set:

- a subset of *tst2015*
  - ~10,000 words
  - ~ first half of the 12 TED talks composing *tst2015*
- *EnDe*: 600 segments
- *ViEn*: 500 segments

Evaluation Setup

Lesson learned from IWSLT 2013/2014:

- most informative and reliable HTER:
  - not by using the targeted reference only
  - but by exploiting all post-edits

Evaluation Setup

Lesson learned from IWSLT 2013/2014:

- most informative and reliable HTER:
  - not by using the targeted reference only
  - but by exploiting all post-edits

SRC:
Tôi lớn lên trong điều kiện nuôi dạy bình thường.

Targeted Reference Only

REF: I had a normal kind of upbringing .
HYP: I **grew** ***up*** ***in*** **[normal]** ***the*** ***conditions*** ***raised*** normal .

TER:
87.50

All Post-Edited References

REF: I grew up in normal raising conditions .
HYP: I grew up in **[normal]** ***the*** conditions **raised** normal .

TER:
38.46

Evaluation Setup

Lesson learned from IWSLT 2013/2014:

- most informative and reliable HTER:
  - not by using the targeted reference only
  - but by exploiting all post-edits

IWSLT 2015 official evaluation:

- HTER calculated on multiple references (post-edits)
  - *EnDe:* 5 participants => 5 post-edits
  - *ViEn:* 5 participants => 5 post-edits

Data Collection

- *Bilingual* Post-Editing
  - professional translators were required to post-edit the MT output directly according to the source sentence

Data Collection

- *Bilingual* Post-Editing
  - professional translators were required to post-edit the MT output directly according to the source sentence
- Data preparation:
  - 5 systems post-edited by 5 professional translators
    - each translator must p-edit all the HE set sentences
    - each translator must p-edit each sentence only once
    - each MT system must be equally p-edited by all translators
  - MT outputs dispatched to translators both randomly and satisfying the uniform assignment constraints

Data Collection

- *Bilingual* Post-Editing
  - professional translators were required to post-edit the MT output directly according to the source sentence
- Data preparation:
  - 5 systems post-edited by 5 professional translators
    - each translator must p-edit all the HE set sentences
    - each translator must p-edit each sentence only once
    - each MT system must be equally p-edited by all translators
  - MT outputs dispatched to translators both randomly and satisfying the uniform assignment constraints
- MateCat post-editing interface

Collected Data

- Collected Post-edits
  - 5 new references for each sentence in the HE set

Collected Data

- Collected Post-edits
  - 5 new references for each sentence in the HE set
- Post-editors characteristics:

	En-De				***Vi-En***			
	PE Effort	st-dv	Sys TER	st-dv	PE Effort	st-dv	Sys TER	st-dv
PE 1	22.49	16.44	56.43	20.77	37.14	21.25	61.38	20.96
PE 2	42.68	26.51	55.59	20.82	40.38	20.46	60.34	20.94
PE 3	29.21	22.18	56.00	20.49	44.76	23.57	61.66	21.74
PE 4	27.66	15.50	55.77	21.17	46.39	25.71	61.69	21.59
PE 5	22.19	17.62	56.38	20.85	38.57	26.64	60.14	20.43

Collected Data

- Collected Post-edits
  - 5 new references for each sentence in the HE set
- Post-editors characteristics:

	En-De				***Vi-En***			
	PE Effort	st-dv	Sys TER	st-dv	PE Effort	st-dv	Sys TER	st-dv
PE 1	22.49	16.44	56.43	20.77	37.14	21.25	61.38	20.96
PE 2	42.68	26.51	55.59	20.82	40.38	20.46	60.34	20.94
PE 3	29.21	22.18	56.00	20.49	44.76	23.57	61.66	21.74
PE 4	27.66	15.50	55.77	21.17	46.39	25.71	61.69	21.59
PE 5	22.19	17.62	56.38	20.85	38.57	26.64	60.14	20.43

- PE effort (HTER): highly variable among post-editors

Collected Data

- Collected Post-edits
  - 5 new references for each sentence in the HE set
- Post-editors characteristics:

	En-De				***Vi-En***			
	PE Effort	st-dv	Sys TER	st-dv	PE Effort	st-dv	Sys TER	st-dv
PE 1	22.49	16.44	56.43	20.77	37.14	21.25	61.38	20.96
PE 2	42.68	26.51	55.59	20.82	40.38	20.46	60.34	20.94
PE 3	29.21	22.18	56.00	20.49	44.76	23.57	61.66	21.74
PE 4	27.66	15.50	55.77	21.17	46.39	25.71	61.69	21.59
PE 5	22.19	17.62	56.38	20.85	38.57	26.64	60.14	20.43

- PE effort (HTER): highly variable among post-editors
- MT outputs assigned to translators (Sys TER): very homogeneous

Evaluation Results - *EnDe*

System	**HTER**	**HTER**	**TER**	**TER**
Ranking	**HE Set**	**HE Set**	**HE Set**	**Test Set**
	All PErefs	**tgt PEref**	**ref**	**ref**
SU	16.16	21.09	51.15	51.13
UEDIN	21.84	27.99	56.39	56.05
KIT	22.67	28.98	55.82	55.52
HDU	23.42	29.93	57.32	56.94
PJAIT	28.18	35.68	59.51	59.03

Rank corr.		1.00	0.90	0.90

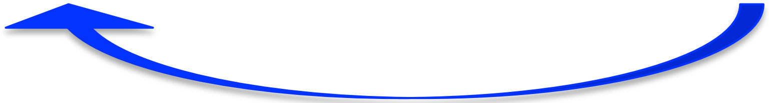
Evaluation Results - *EnDe*

System Ranking	HTER HE Set All PErefs	HTER HE Set tgt PEref	TER HE Set ref	TER Test Set ref
SU	16.16	21.09	51.15	51.13
UEDIN	21.84	27.99	56.39	56.05
KIT	22.67	28.98	55.82	55.52
HDU	23.42	29.93	57.32	56.94
PJAIT	28.18	35.68	59.51	59.03
Rank corr.	↑	1.00	0.90	0.90

Statistical Significance at *p* < 0.01
(Approximate Randomization)

Evaluation Results - *EnDe*

System Ranking	HTER HE Set All PErefs	HTER HE Set tgt PEref	TER HE Set ref	TER Test Set ref
SU	16.16	21.09	51.15	51.13
UEDIN	21.84	27.99	56.39	56.05
KIT	22.67	28.98	55.82	55.52
HDU	23.42	29.93	57.32	56.94
PJAIT	28.18	35.68	59.51	59.03
Rank corr.		1.00	0.90	0.90



TER/HTER reduction

Evaluation Results - *EnDe*

System Ranking	**HTER HE Set All PErefs**	**HTER HE Set tgt PEref**	**TER HE Set ref**	**TER Test Set ref**
SU	16.16	21.09	51.15	51.13
UEDIN	21.84	27.99	56.39	56.05
KIT	22.67	28.98	55.82	55.52
HDU	23.42	29.93	57.32	56.94
PJAIT	28.18	35.68	59.51	59.03
Rank corr.		1.00	0.90	0.90

↑

Spearman’s Rank Coefficient

Evaluation Results - Vi*En*

System Ranking	**HTER HE Set All PErefs**	**HTER HE Set tgt PEref**	**TER HE Set ref**	**TER Test Set ref**
JAIST	32.24	37.25	60.10	62.35
UMD	32.71	37.99	58.92	59.19
PJAIT	34.27	40.50	59.48	62.20
TUT	38.50	43.42	62.49	62.69
UNETI	41.42	47.97	64.21	66.33

Rank corr.		1.00	0.70	0.70

Evaluation Results - Vi*En*

System	HTER	HTER	TER	TER
Ranking	HE Set	HE Set	HE Set	Test Set
	All PErefs	tgt PEref	ref	ref
JAIST	32.24	37.25	60.10	62.35
UMD	32.71	37.99	58.92	59.19
PJAIT	34.27*	40.50	59.48	62.20
TUT	38.50	43.42	62.49	62.69
UNETI	41.42	47.97	64.21	66.33
Rank corr.		1.00	0.70	0.70

Statistical Significance at $p < 0.01$ (* = $p < 0.05$)
(Approximate Randomization)

Evaluation Results - Vi*En*

System Ranking	**HTER HE Set All PErefs**	**HTER HE Set tgt PEref**	**TER HE Set ref**	**TER Test Set ref**
JAIST	32.24	37.25	60.10	62.35
UMD	32.71	37.99	58.92	59.19
PJAIT	34.27	40.50	59.48	62.20
TUT	38.50	43.42	62.49	62.69
UNETI	41.42	47.97	64.21	66.33
Rank corr.		1.00	0.70	0.70

TER/HTER reduction

Evaluation Results - Vi*En*

System Ranking	**HTER HE Set All PErefs**	**HTER HE Set tgt PEref**	**TER HE Set ref**	**TER Test Set ref**
JAIST	32.24	37.25	60.10	62.35
UMD	32.71	37.99	58.92	59.19
PJAIT	34.27	40.50	59.48	62.20
TUT	38.50	43.42	62.49	62.69
UNETI	41.42	47.97	64.21	66.33
Rank corr.		1.00	0.70	0.70

Spearman’s Rank Coefficient

Future

- TED task by now very seasoned
    - Extend to more realistic lectures
    - Work on more challenging tasks: conversations
- Include more under-resourced languages on the input side
- Discussion on co-location with another MT/NLP conference
- Continue with HE based on post-editing
    - Funding by H2020 CSA Cracker

Detailed discussion with proposals for new tasks tomorrow

Credits

- **Language resources**
  - TED LLC, USA (Talk data)
  - Workshop Machine Translation (Giga and news data)
  - DFKI, Germany (United Nations data)
  - PJAIT (Wikipedia parallel corpus)
  - Cantab Reserarch (LM and text corpus for TED)
  - Many other external data providers
- **Funding**
  - H2020 CSA CRACKER
  - Internal funds of eval organizers
  - …