The IWSLT 2016 Evaluation Campaign

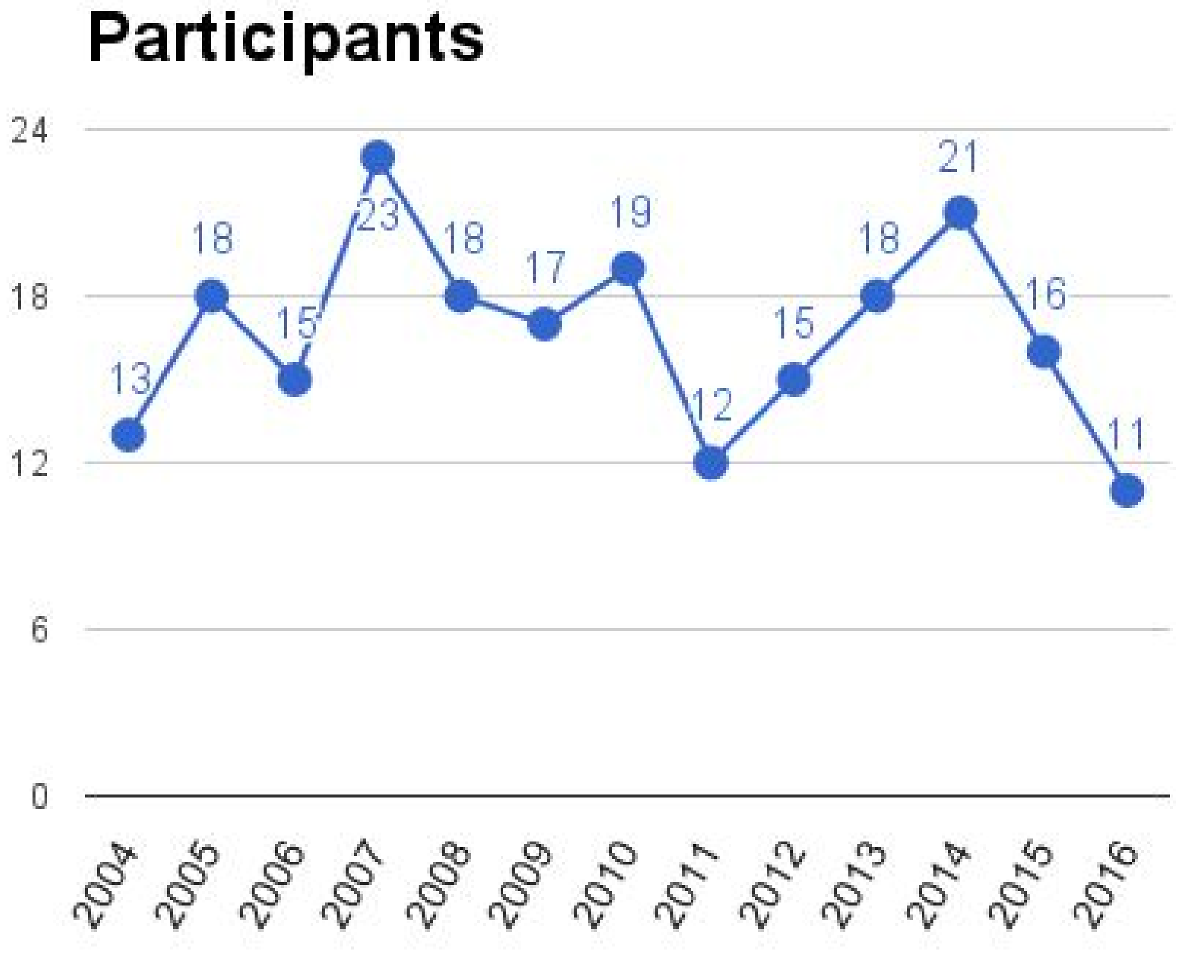
Mauro Cettolo, FBK, Italy
Jan Niehues, KIT, Germany
Sebastian Stüker, KIT, Germany
Luisa Bentivogli, FBK, Italy
Roldano Cattoni, FBK, Italy
Marcello Federico, FBK, Italy

Redmond, WA, 8-9 December 2016

Outline

- IWSLT review
- Tasks and Tracks
- Participants
- Automatic evaluation
- Human evaluation
- Future plans

IWSLT Evaluation: record of participants



Tasks

- Talks Task: SLT for subtitling
  - TED data
  - QED data
- MSLT Task: SLT for video conference
  - MSLT data

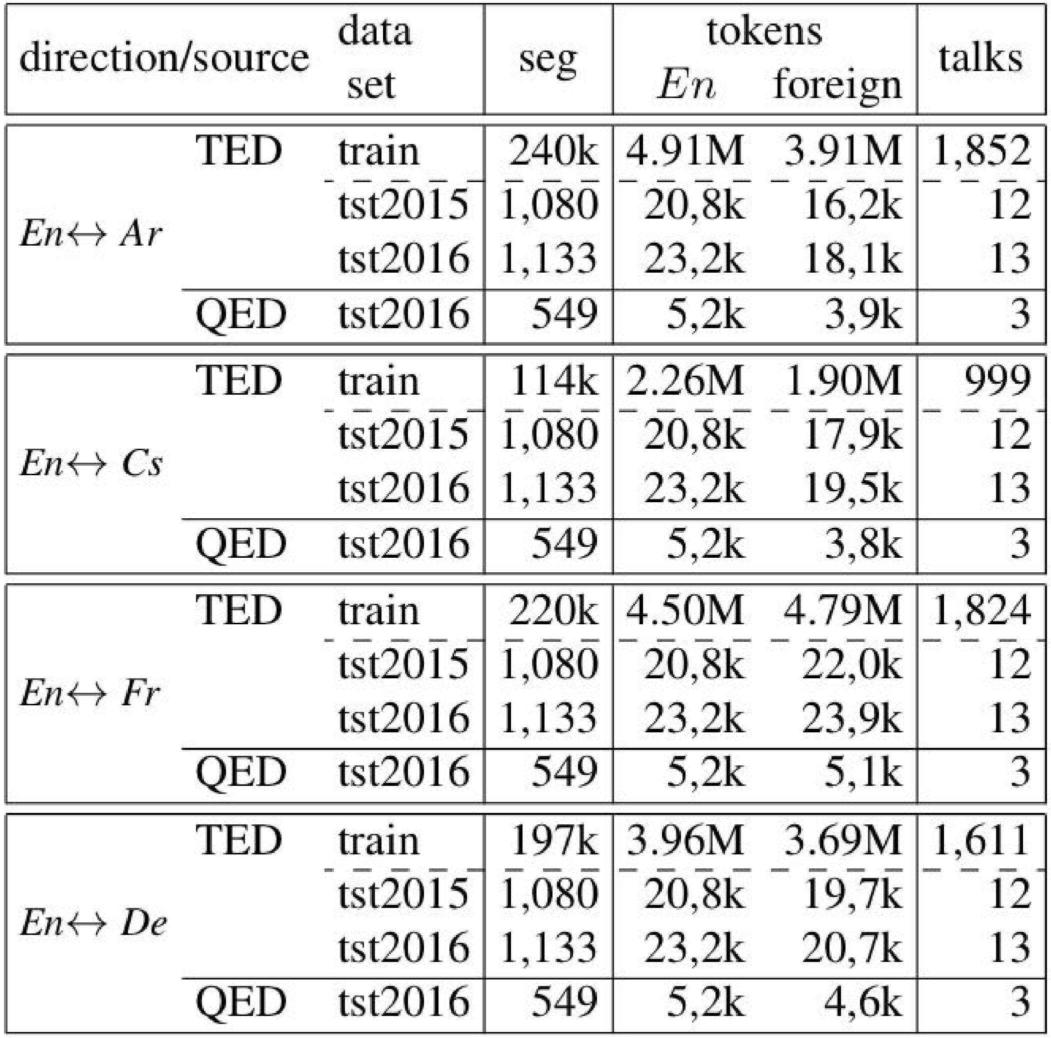
Talks Task

English > French

English > German

(*) Fr|De >En, En <> Ar|Cz only from text

Talk Task Resources



direction/source		data set	seg	tokens		talks
				En	foreign	
En↔ *Ar*	TED	train	240k	4.91M	3.91M	1,852
		tst2015	1,080	20,8k	16,2k	12
		tst2016	1,133	23,2k	18,1k	13
	QED	tst2016	549	5,2k	3,9k	3
En↔ *Cs*	TED	train	114k	2.26M	1.90M	999
		tst2015	1,080	20,8k	17,9k	12
		tst2016	1,133	23,2k	19,5k	13
	QED	tst2016	549	5,2k	3,8k	3
En↔ *Fr*	TED	train	220k	4.50M	4.79M	1,824
		tst2015	1,080	20,8k	22,0k	12
		tst2016	1,133	23,2k	23,9k	13
	QED	tst2016	549	5,2k	5,1k	3
En↔ *De*	TED	train	197k	3.96M	3.69M	1,611
		tst2015	1,080	20,8k	19,7k	12
		tst2016	1,133	23,2k	20,7k	13
	QED	tst2016	549	5,2k	4,6k	3

QED corpus site contains IWSLT 2016 distribution!

Challenges in Talk Task

Language/translation modeling

- ➢Variability of topics and styles
- ➢Distant languages, morphology

Audio/speech modeling

- ➢Noise: mumble, applauses, laughs, music, ...
- ➢Speaker: accent, speaking rate, style,
  spontaneous speech phenomena (esp. on QED)

MSLT Task

English <> German

English > French

MSLT Task dataset

Transcript (w/ disfluencies):

ähm wir haben grade über Platten geredet, und über, über Musik, Musik Stream, was mich halt irgendwie nervt ist das bei so vielen Platten vorn so krass viel Werbung dazwischen geschaltet wird, und das find ich äh sehr störend, ja.

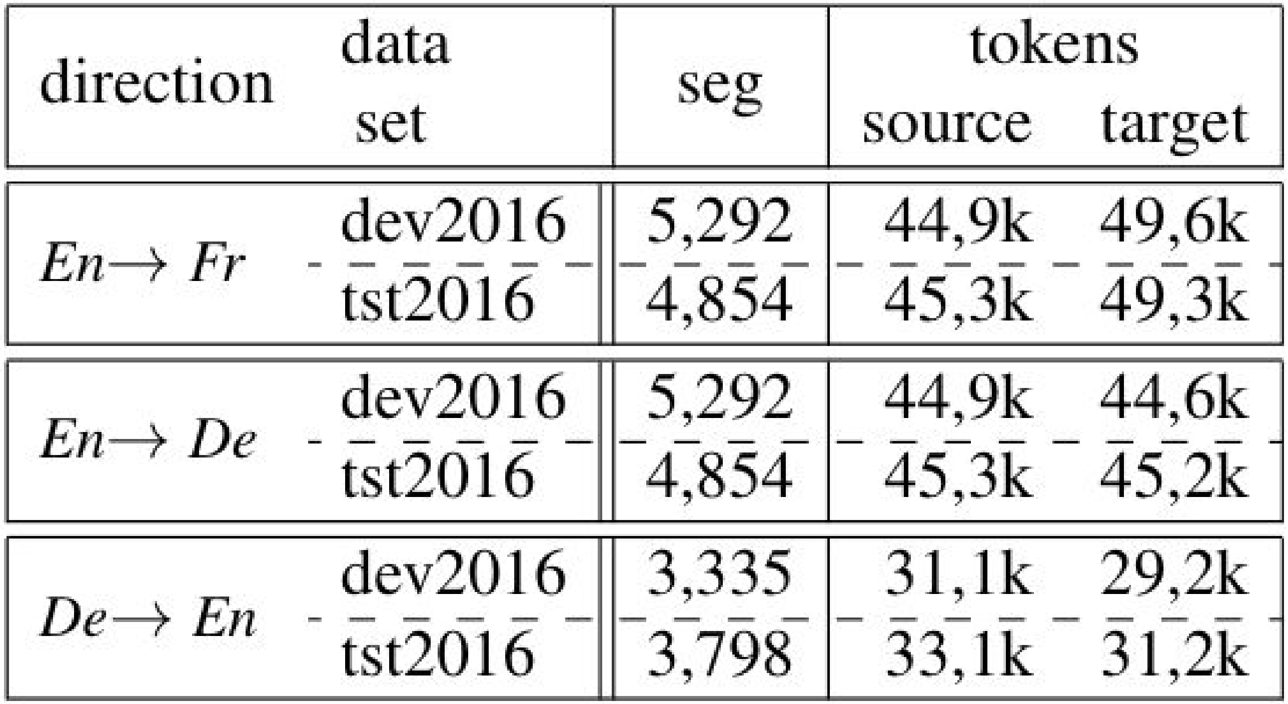
Polished text (w/o disfluences):

Wir haben grade über Platten geredet und über Musik Stream, was mich halt irgendwie nervt ist, dass bei so vielen Platten vorn so krass viel Werbung dazwischen geschaltet wird. Und das find ich sehr störend, ja.

Translation into English:

We just talked about albums and about streaming music, which just bugs me somehow, that for so many albums, so much advertising is placed before and in between them. And I find that very disruptive, yes.

MSLT Task Resources



direction	data set	seg	tokens	
			source	target
En→ *Fr*	dev2016	5,292	44,9k	49,6k
	tst2016	4,854	45,3k	49,3k
En→ *De*	dev2016	5,292	44,9k	44,6k
	tst2016	4,854	45,3k	45,2k
De→ *En*	dev2016	3,335	31,1k	29,2k
	tst2016	3,798	33,1k	31,2k

No task specific training data available

Challenges in MSLT Task

Language/translation modelling

- ➢No task-specific training data
- ➢Word order, morphology
- ➢ Conversational speech

Acoustic modelling

- ➢Noise: channel
- ➢Speaker: disfluencies, code switching, ...

2016 Tracks

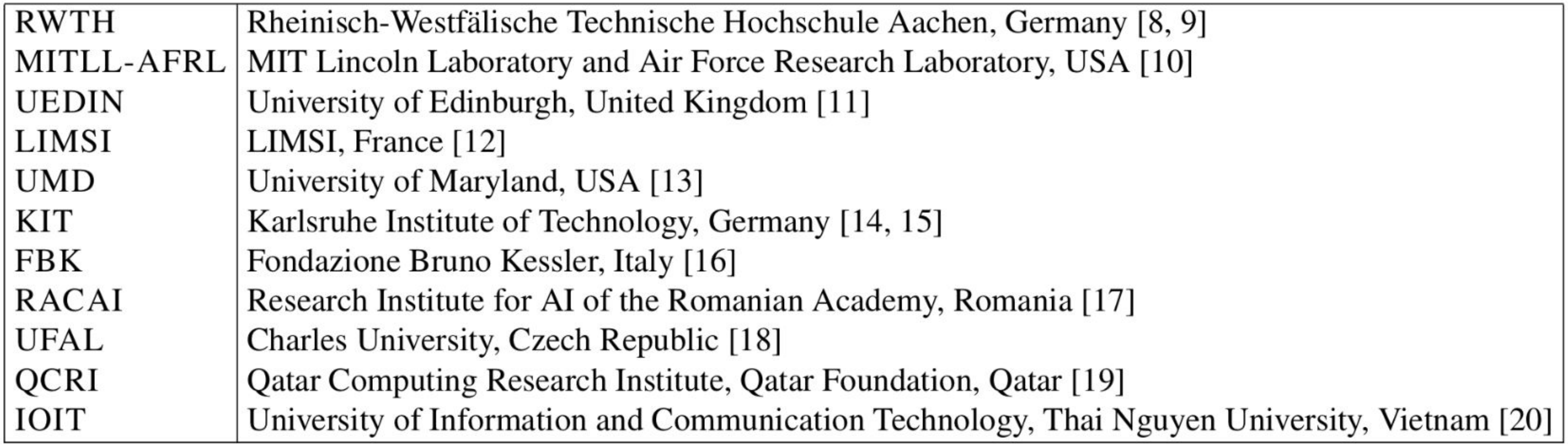
- **Automatic Speech Recognition (ASR)**
  - Transcription from audio to text
  - English (TALK,MSLT), German (MSLT)
- **Spoken Language Translation (SLT)**
  - Translation from audio (or ASR output) to text
  - English > German, French (TALK)
  - English <> German, English > French (MSLT)
- **Machine Translation (MT)**
  - Translation from text (cleaned transcripts) to text (translation)
  - English <> German, French, Czech, Arabic (TALK)
  - English <> German, English>French (MSLT)

Specifications

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

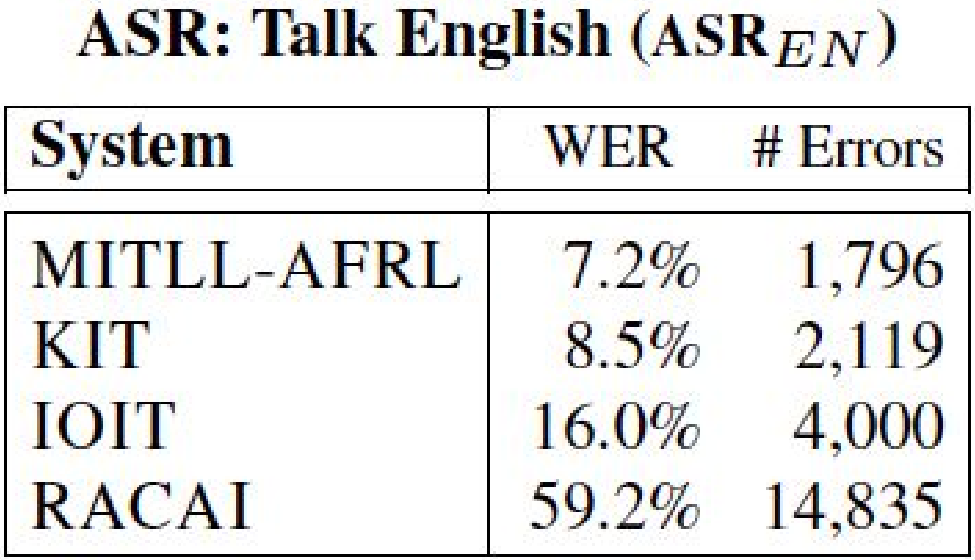
Metrics	ASR	SLT	MT
WER	✔		
BLEU		✔	✔
TER		✔	✔
NIST			✔

Participants



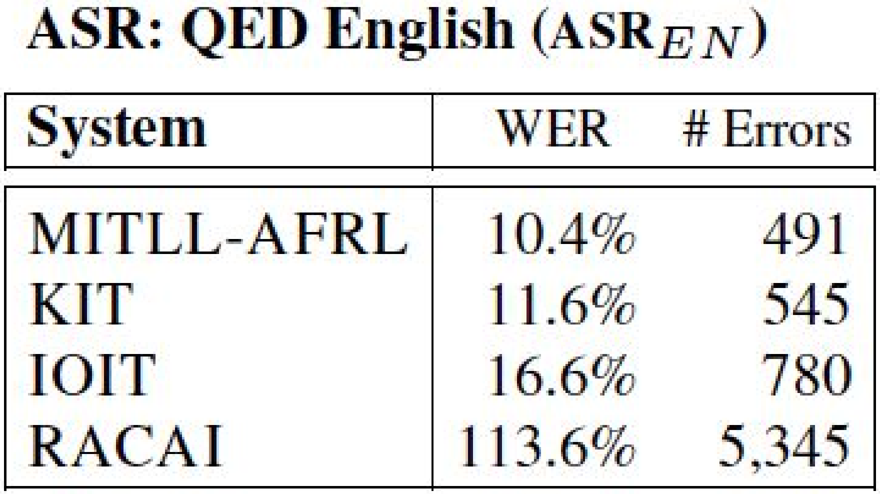
RWTH	Rheinisch-Westfälische Technische Hochschule Aachen, Germany [8, 9]
MITLL-AFRL	MIT Lincoln Laboratory and Air Force Research Laboratory, USA [10]
UEDIN	University of Edinburgh, United Kingdom [11]
LIMSI	LIMSI, France [12]
UMD	University of Maryland, USA [13]
KIT	Karlsruhe Institute of Technology, Germany [14, 15]
FBK	Fondazione Bruno Kessler, Italy [16]
RACAI	Research Institute for AI of the Romanian Academy, Romania [17]
UFAL	Charles University, Czech Republic [18]
QCRI	Qatar Computing Research Institute, Qatar Foundation, Qatar [19]
IOIT	University of Information and Communication Technology, Thai Nguyen University, Vietnam [20]

Results: ASR



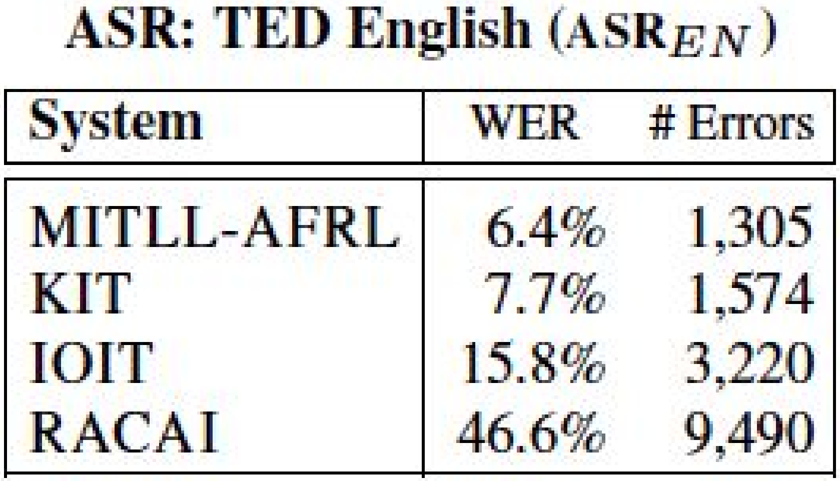
ASR: Talk English (ASR_{EN})

System	WER	# Errors
MITLL-AFRL	7.2%	1,796
KIT	8.5%	2,119
IOIT	16.0%	4,000
RACAI	59.2%	14,835



ASR: QED English (ASR_{EN})

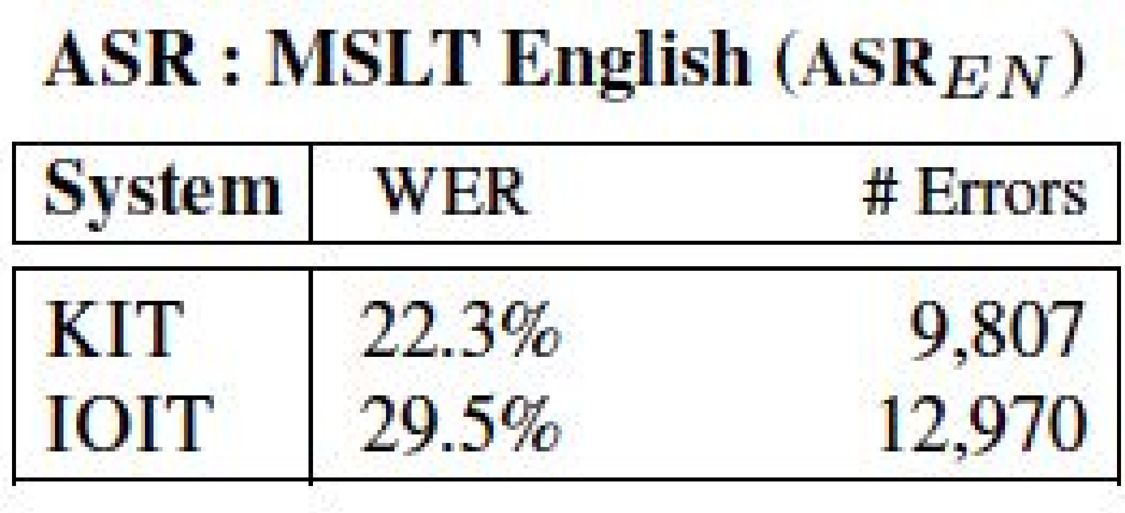
System	WER	# Errors
MITLL-AFRL	10.4%	491
KIT	11.6%	545
IOIT	16.6%	780
RACAI	113.6%	5,345



ASR: TED English (ASR_{EN})

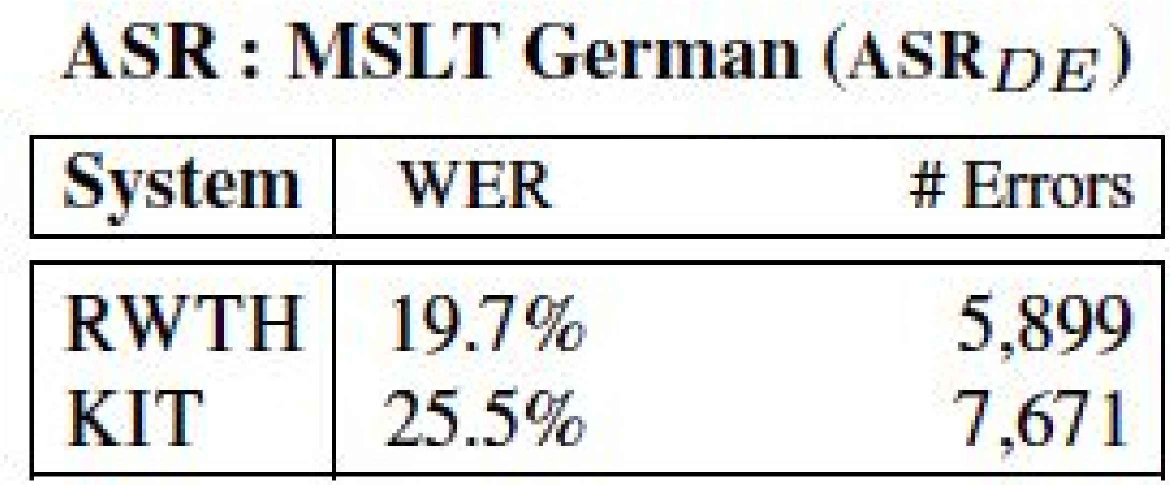
System	WER	# Errors
MITLL-AFRL	6.4%	1,305
KIT	7.7%	1,574
IOIT	15.8%	3,220
RACAI	46.6%	9,490

Results: ASR



ASR : MSLT English (ASR_{EN})

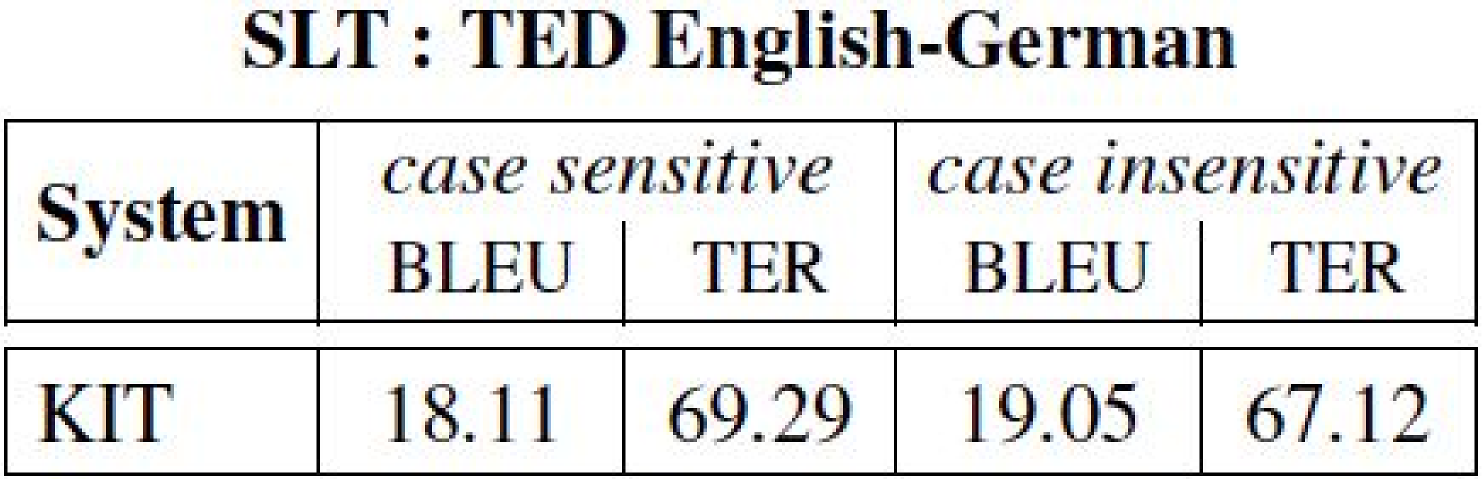
System	WER	# Errors
KIT	22.3%	9,807
IOIT	29.5%	12,970



ASR : MSLT German (ASR_{DE})

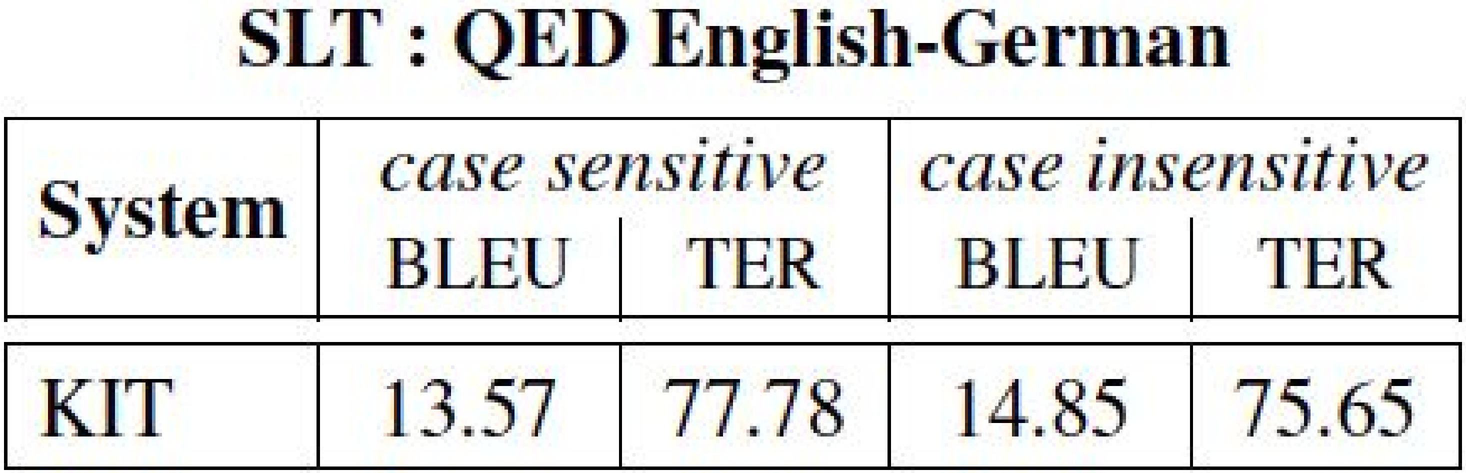
System	WER	# Errors
RWTH	19.7%	5,899
KIT	25.5%	7,671

Results: SLT



SLT : TED English-German

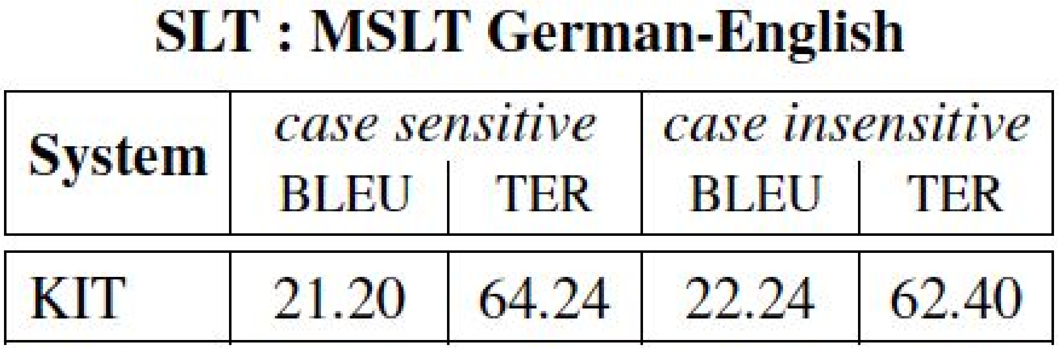
System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	18.11	69.29	19.05	67.12



SLT : QED English-German

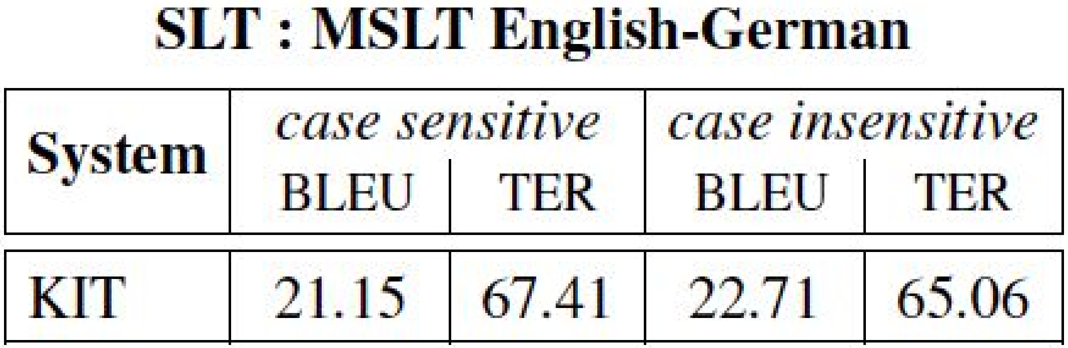
System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	13.57	77.78	14.85	75.65

Results: SLT



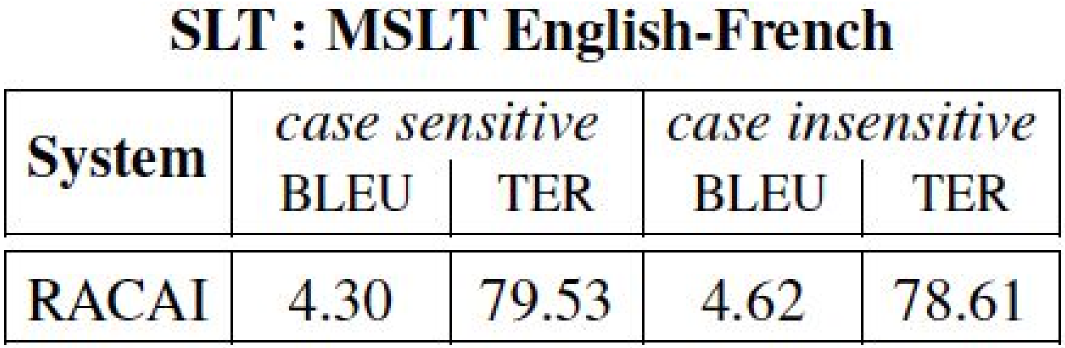
SLT : MSLT German-English

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	21.20	64.24	22.24	62.40



SLT : MSLT English-German

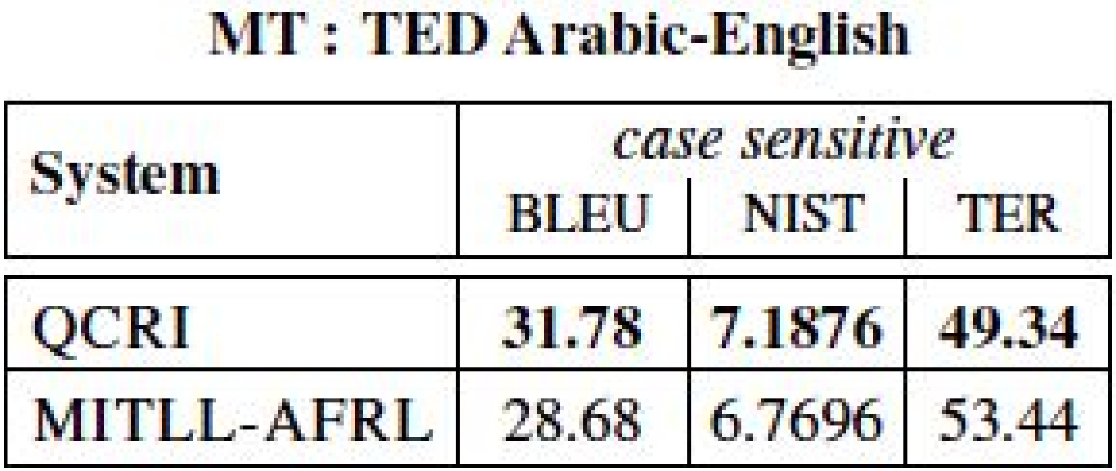
System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	21.15	67.41	22.71	65.06



SLT : MSLT English-French

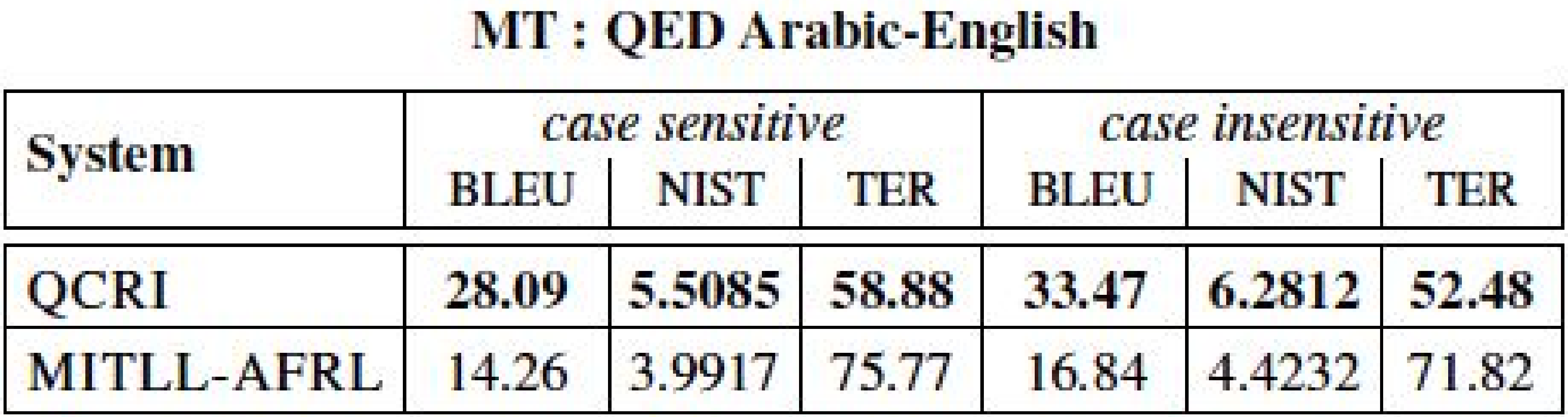
System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
RACAI	4.30	79.53	4.62	78.61

Results: MT



MT : TED Arabic-English

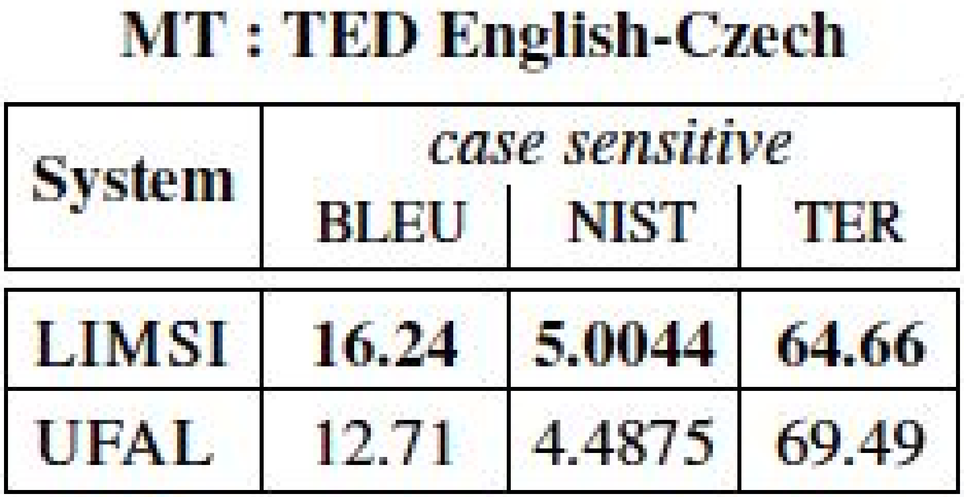
System	*case sensitive*		
	BLEU	NIST	TER
QCRI	**31.78**	**7.1876**	**49.34**
MITLL-AFRL	28.68	6.7696	53.44



MT : QED Arabic-English

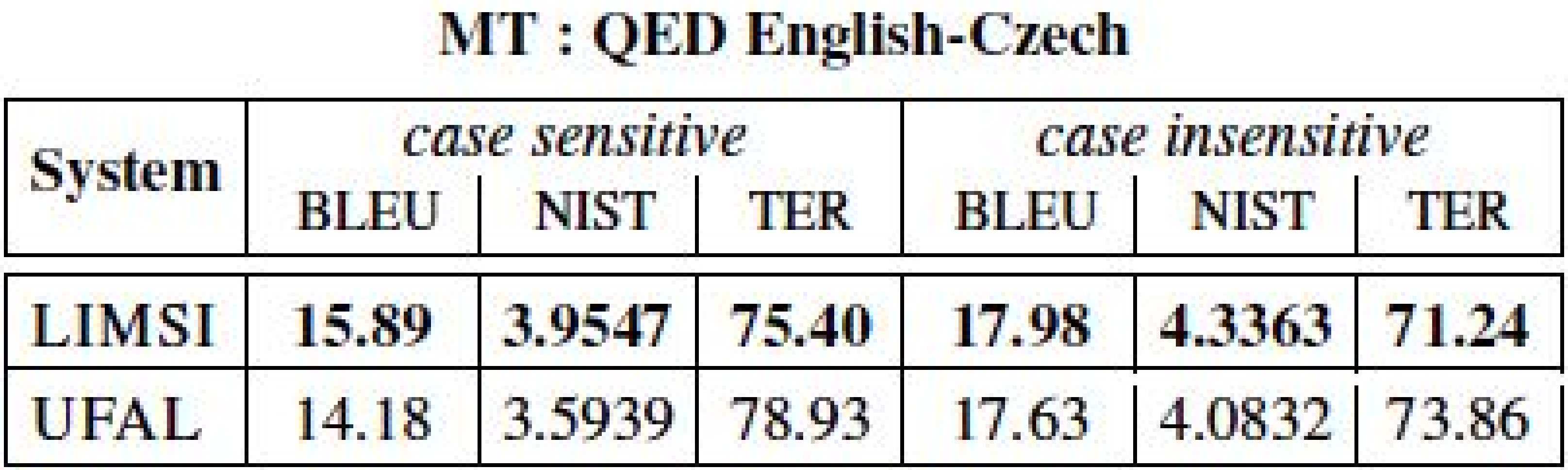
System	*case sensitive*			*case insensitive*		
	BLEU	NIST	TER	BLEU	NIST	TER
QCRI	**28.09**	**5.5085**	**58.88**	**33.47**	**6.2812**	**52.48**
MITLL-AFRL	14.26	3.9917	75.77	16.84	4.4232	71.82

Results: MT



MT : TED English-Czech

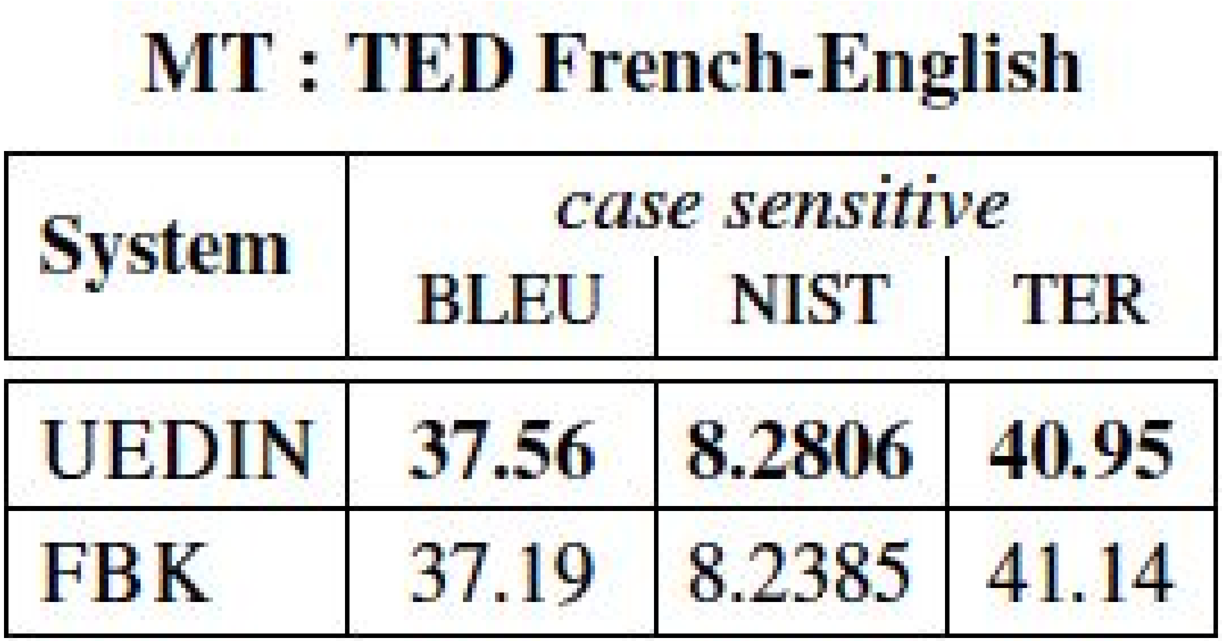
System	*case sensitive*		
	BLEU	NIST	TER
LIMSI	**16.24**	**5.0044**	**64.66**
UFAL	12.71	4.4875	69.49



MT : QED English-Czech

System	*case sensitive*			*case insensitive*		
	BLEU	NIST	TER	BLEU	NIST	TER
LIMSI	**15.89**	**3.9547**	**75.40**	**17.98**	**4.3363**	**71.24**
UFAL	14.18	3.5939	78.93	17.63	4.0832	73.86

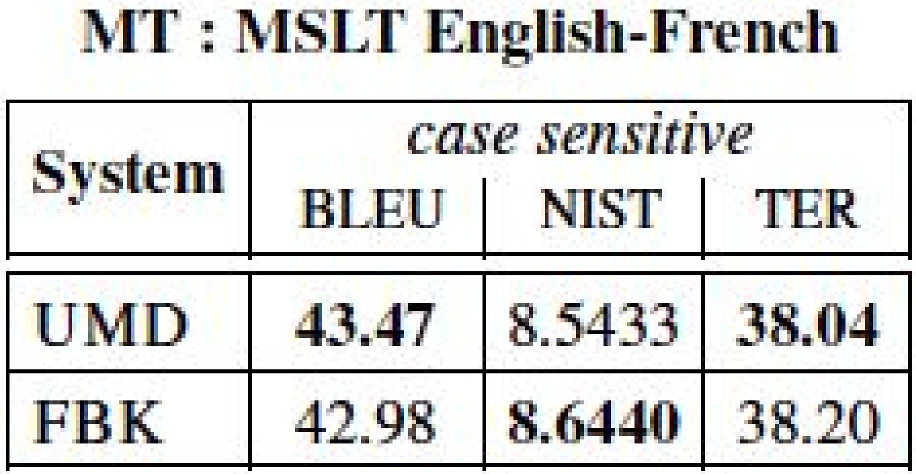
Results: MT



MT : TED French-English

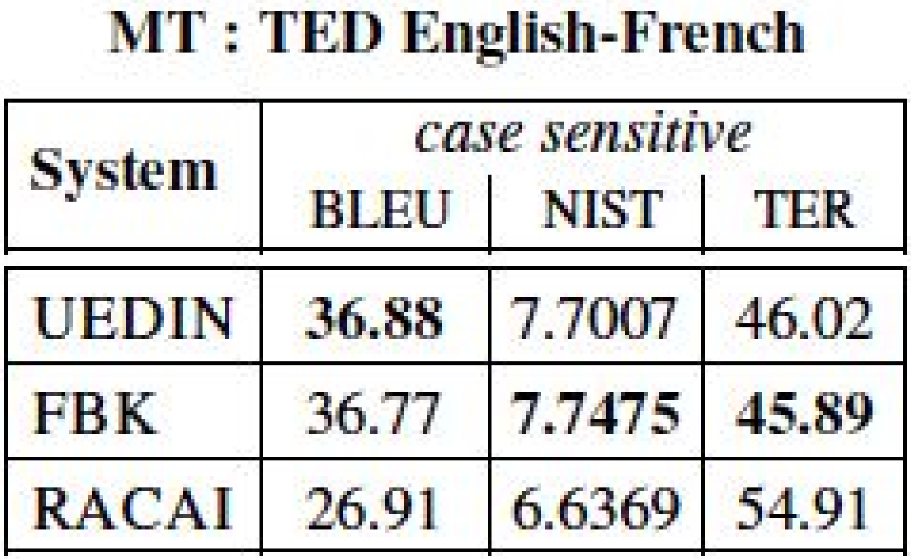
System	*case sensitive*		
	BLEU	NIST	TER
UEDIN	**37.56**	**8.2806**	**40.95**
FBK	37.19	8.2385	41.14

Results: MT



MT : MSLT English-French

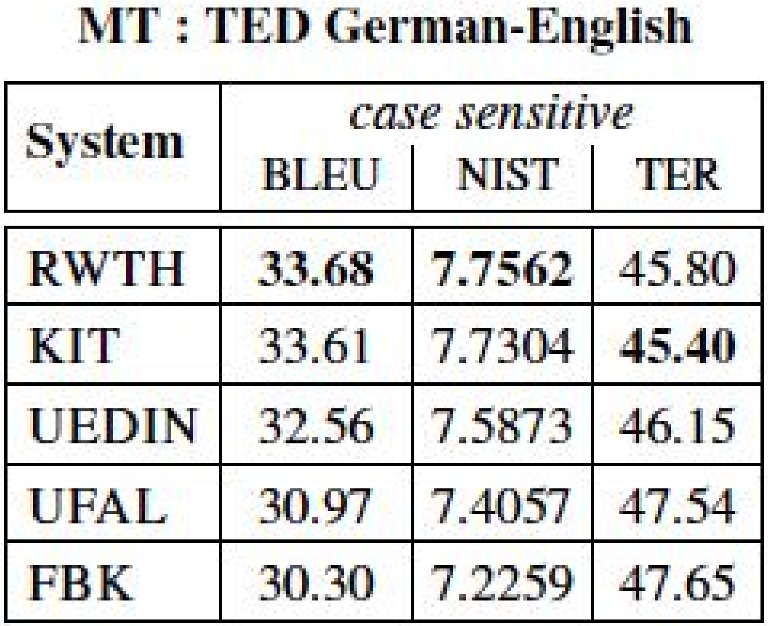
System	*case sensitive*		
	BLEU	NIST	TER
UMD	**43.47**	8.5433	**38.04**
FBK	42.98	**8.6440**	38.20



MT : TED English-French

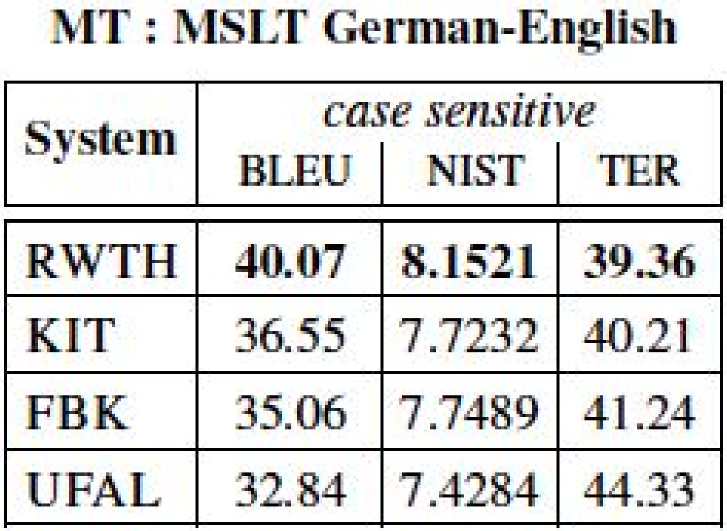
System	*case sensitive*		
	BLEU	NIST	TER
UEDIN	**36.88**	7.7007	46.02
FBK	36.77	**7.7475**	**45.89**
RACAI	26.91	6.6369	54.91

Results: MT



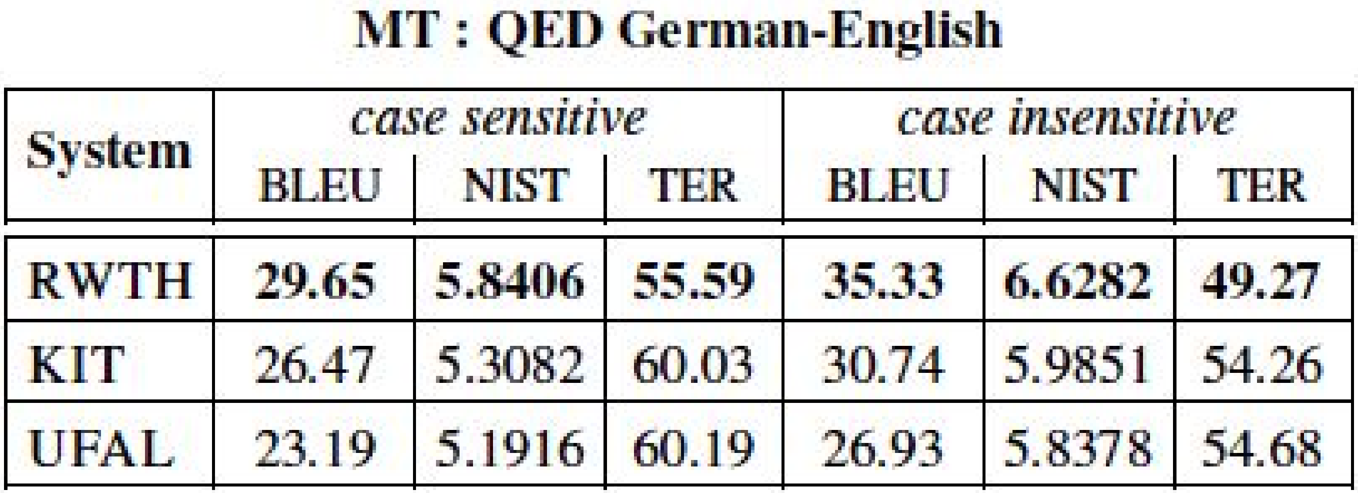
MT : TED German-English

System	*case sensitive*		
	BLEU	NIST	TER
RWTH	**33.68**	**7.7562**	45.80
KIT	33.61	7.7304	**45.40**
UEDIN	32.56	7.5873	46.15
UFAL	30.97	7.4057	47.54
FBK	30.30	7.2259	47.65



MT : MSLT German-English

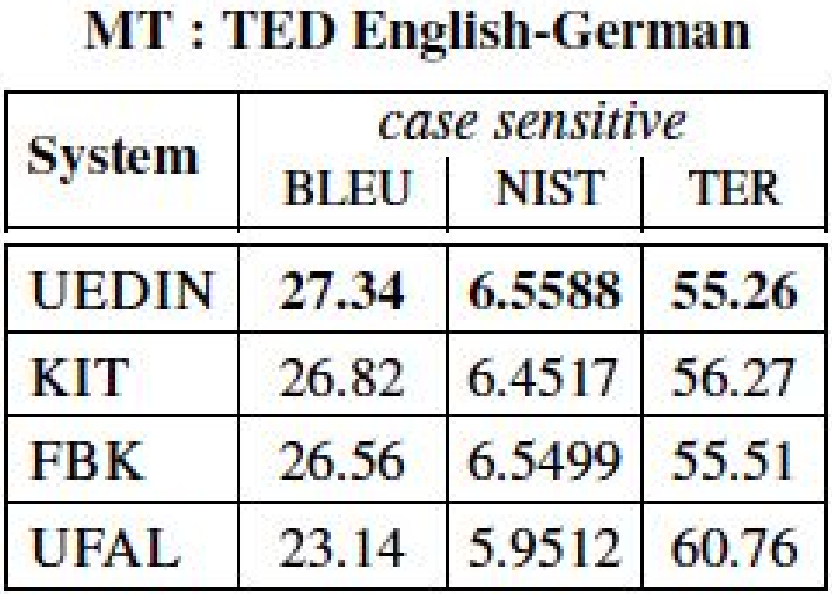
System	*case sensitive*		
	BLEU	NIST	TER
RWTH	**40.07**	**8.1521**	**39.36**
KIT	36.55	7.7232	40.21
FBK	35.06	7.7489	41.24
UFAL	32.84	7.4284	44.33



MT : QED German-English

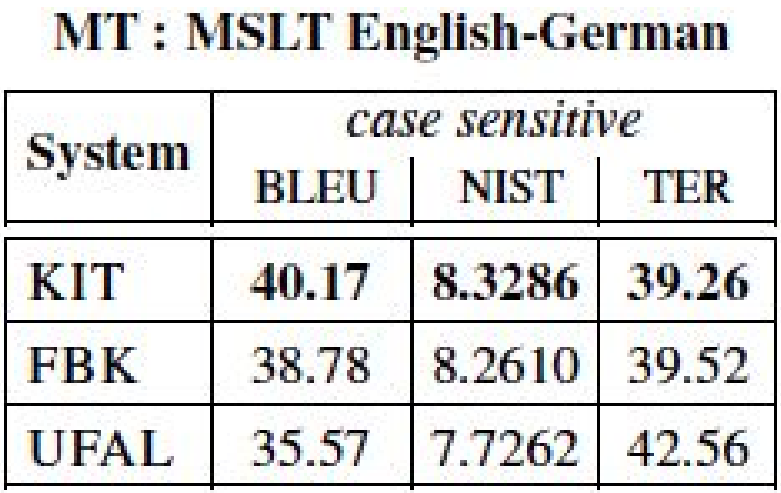
System	*case sensitive*			*case insensitive*		
	BLEU	NIST	TER	BLEU	NIST	TER
RWTH	**29.65**	**5.8406**	**55.59**	**35.33**	**6.6282**	**49.27**
KIT	26.47	5.3082	60.03	30.74	5.9851	54.26
UFAL	23.19	5.1916	60.19	26.93	5.8378	54.68

Results: MT



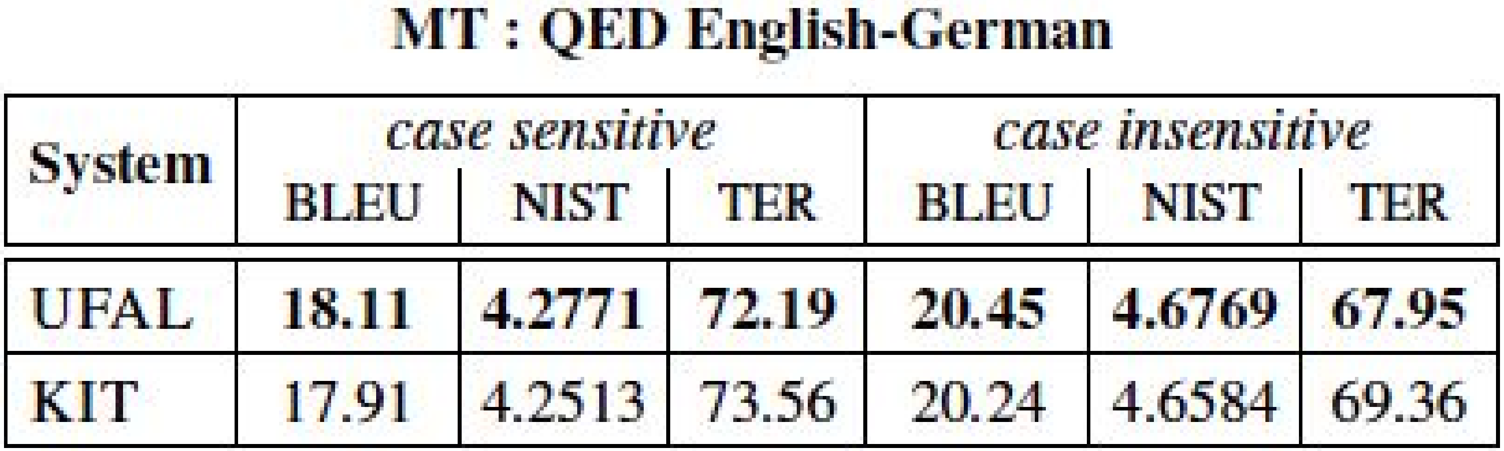
MT : TED English-German

System	*case sensitive*		
	BLEU	NIST	TER
UEDIN	**27.34**	**6.5588**	**55.26**
KIT	26.82	6.4517	56.27
FBK	26.56	6.5499	55.51
UFAL	23.14	5.9512	60.76



MT : MSLT English-German

System	*case sensitive*		
	BLEU	NIST	TER
KIT	**40.17**	**8.3286**	**39.26**
FBK	38.78	8.2610	39.52
UFAL	35.57	7.7262	42.56



MT : QED English-German

System	*case sensitive*			*case insensitive*		
	BLEU	NIST	TER	BLEU	NIST	TER
UFAL	**18.11**	**4.2771**	**72.19**	**20.45**	**4.6769**	**67.95**
KIT	17.91	4.2513	73.56	20.24	4.6584	69.36

Human Evaluation

- Following IWSLT 2013/14/15: ***Post-Editing + TER***
  - TED task as an interesting application scenario to test the utility of MT systems in a real subtitling task
  - Edits point to specific translation errors
  - TER traces the edits done by post-editors
  - Additional reference translations
- Evaluation of **MT-*EnDe*** and **MT-*EnFr*** tasks
- Performed on 2015 test set (*tst2015*)

Evaluation Data

tst 2015 HE SET

12 TED Talks

- initial 56% of each talk
- 600 src sentences
- ~10K src words

Evaluation Data

same dataset for EnDe and EnFr

tst 2015 HE SET

12 TED Talks

- initial 56% of each talk
- 600 src sentences
- ~10K src words

Evaluation Data

tst 2015 HE SET

12 TED Talks

- initial 56% of each talk
- 600 src sentences
- ~10K src words

SYS-1

SYS-2

SYS-3

SYS-*n*

Evaluation Data

tst 2015 HE SET

12 TED Talks

- initial 56% of each talk
- 600 src sentences
- ~10K src words

EnDe: 4 systems

EnFr: 5 systems

SYS-1

SYS-2

SYS-3

SYS-*n*

Evaluation Data

tst 2015 HE SET

12 TED Talks

- initial 56% of each talk
- 600 src sentences
- ~10K src words

SYS-1

SYS-2

SYS-3

SYS-*n*

SYS-1 Post-Edit

SYS-2 Post-Edit

SYS-3 Post-Edit

SYS-*n* Post-Edit

Evaluation Data

tst 2015 HE SET

12 TED Talks

- initial 56% of each talk
- 600 src sentences
- ~10K src words

SYS-1

SYS-2

SYS-3

SYS-*n*

SYS-1 Post-Edit

SYS-2 Post-Edit

SYS-3 Post-Edit

SYS-*n* Post-Edit

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

Evaluation Data

tst 2015 HE SET

12 TED Talks

- initial 56% of each talk
- 600 src sentences
- ~10K src words

SYS-1

SYS-1 Post-Edit

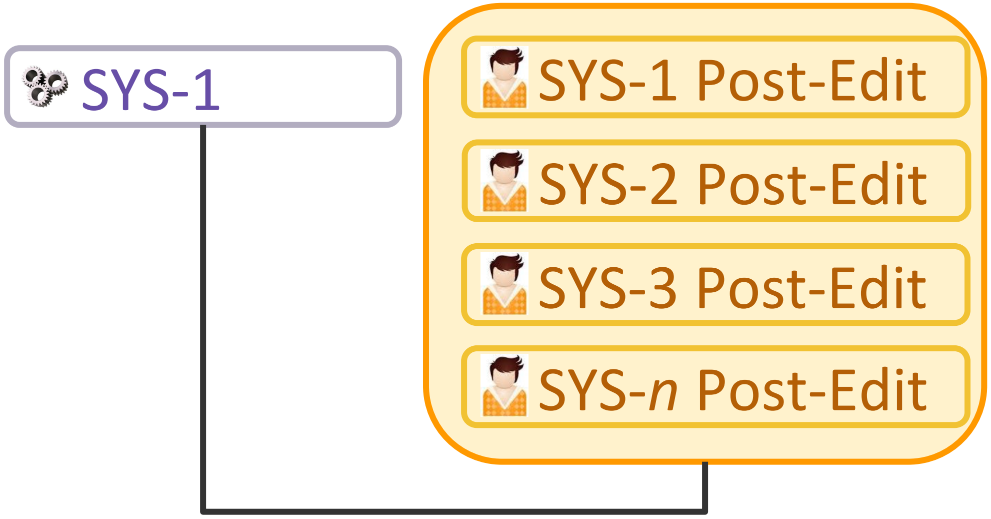
Targeted post-edit
(HTER)

Evaluation Data

tst 2015 HE SET

12 TED Talks

- initial 56% of each talk
- 600 src sentences
- ~10K src words



SYS-1

SYS-1 Post-Edit

SYS-2 Post-Edit

SYS-3 Post-Edit

SYS-*n* Post-Edit

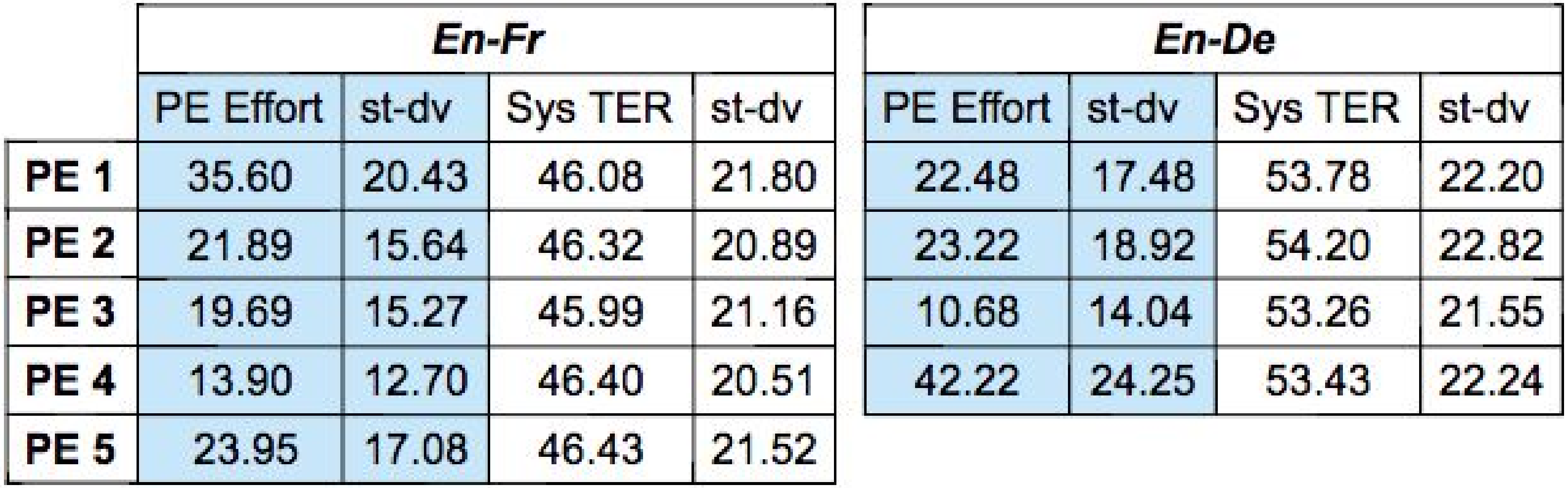
Multiple references
(mTER)

Post-editor analysis

Post-Editing effort:

- the number of actual edit operations performed to produce the post-edited version
- calculated with HTER

➔ highly variable among post-editors



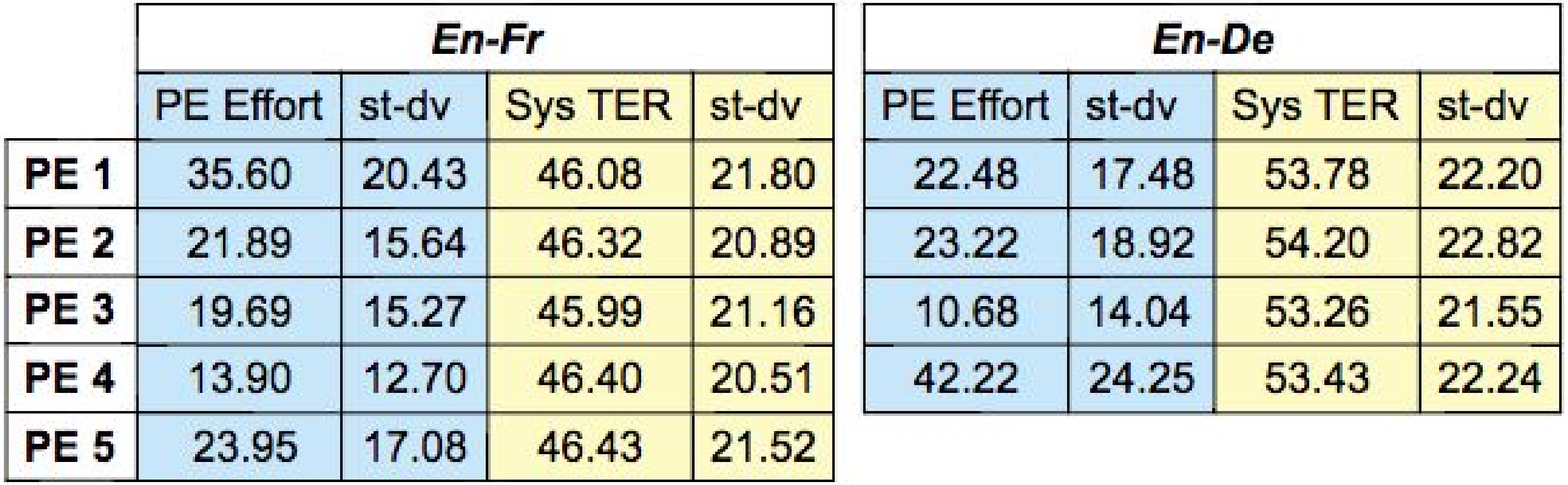
	En-Fr				*En-De*			
	PE Effort	st-dv	Sys TER	st-dv	PE Effort	st-dv	Sys TER	st-dv
PE 1	35.60	20.43	46.08	21.80	22.48	17.48	53.78	22.20
PE 2	21.89	15.64	46.32	20.89	23.22	18.92	54.20	22.82
PE 3	19.69	15.27	45.99	21.16	10.68	14.04	53.26	21.55
PE 4	13.90	12.70	46.40	20.51	42.22	24.25	53.43	22.24
PE 5	23.95	17.08	46.43	21.52				

Post-editor analysis

MT outputs assigned to translators:

- calculated with TER against the official reference

➔ very homogeneous



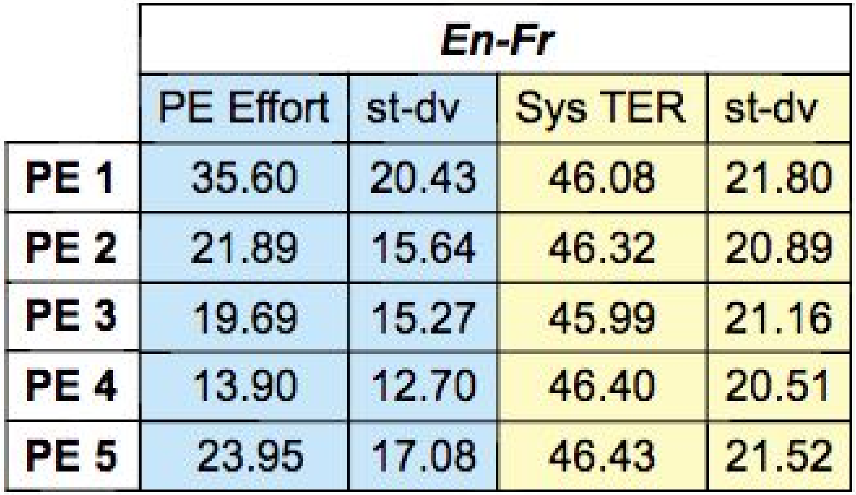
	En-Fr				*En-De*			
	PE Effort	st-dv	Sys TER	st-dv	PE Effort	st-dv	Sys TER	st-dv
PE 1	35.60	20.43	46.08	21.80	22.48	17.48	53.78	22.20
PE 2	21.89	15.64	46.32	20.89	23.22	18.92	54.20	22.82
PE 3	19.69	15.27	45.99	21.16	10.68	14.04	53.26	21.55
PE 4	13.90	12.70	46.40	20.51	42.22	24.25	53.43	22.24
PE 5	23.95	17.08	46.43	21.52				

Post-editor analysis

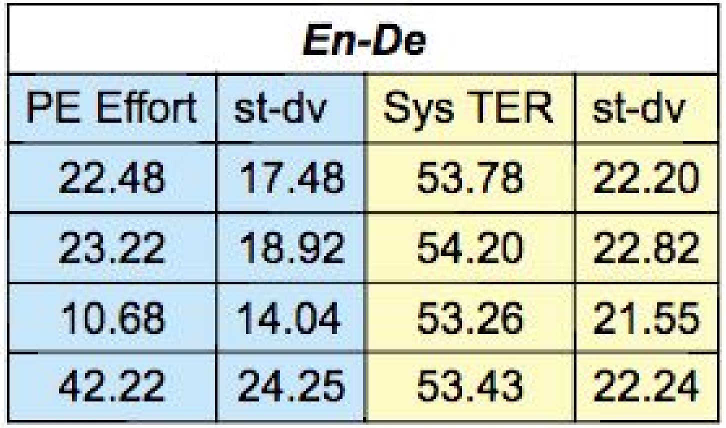
MT outputs assigned to translators:

- calculated with TER against the official reference

➔ very homogeneous



	En-Fr			
	PE Effort	st-dv	Sys TER	st-dv
PE 1	35.60	20.43	46.08	21.80
PE 2	21.89	15.64	46.32	20.89
PE 3	19.69	15.27	45.99	21.16
PE 4	13.90	12.70	46.40	20.51
PE 5	23.95	17.08	46.43	21.52



En-De			
PE Effort	st-dv	Sys TER	st-dv
22.48	17.48	53.78	22.20
23.22	18.92	54.20	22.82
10.68	14.04	53.26	21.55
42.22	24.25	53.43	22.24

Difference due to translators’ subjectivity

Evaluation Metrics

Lesson learned from past IWSLT evaluations

- ➢Most informative assessment of overall MT performance:
  - ➢Not by using the targeted reference only (HTER)
  - ➢But by exploiting all post-edits (mTER)

Evaluation Metrics

Lesson learned from past IWSLT evaluations

- ➢Most informative assessment of overall MT performance:
  - ➢Not by using the targeted reference only (HTER)
  - ➢But by exploiting all post-edits (mTER)

SRC:
But why would you reconcile after a fight?

Targeted Reference Only

HTER: 50.00

REF: Mais pourquoi voudriez-vous **vous réconcilier** après **vous être battu** ?
HYP: Mais pourquoi voudriez-vous **** **concilier** après **** **un combat** ?

All Post-Edits

mTER: 23.33

REF: Mais pourquoi **se** **réconcilier** après un combat ?
HYP: Mais pourquoi **voudriez-vous concilier** après un combat ?

Evaluated Systems

EnDe Task:

- ➢4 submitted primary runs (3 NMT + 1 PBMT)
- ➢1 winning system of IWSLT 2015 (NMT, Stanford)

EnFr Task:

- ➢2 top-ranking primary runs (NMT)
- ➢2 external sota PBMT (GT and ModernMT)
- ➢1 primary submission from IWSLT 2015 (PBMT)

Evaluation Results - *EnDe*

System Ranking	**mTER HE Set 5 PErefs**
UEDIN	13.31
KIT	14.12
SU-2015	14.98
FBK	15.95
UFAL	21.89

Evaluation Results - *EnDe*

System Ranking	mTER HE Set 5 PErefs
UEDIN 🟩	13.31
KIT 🟧 🟩	14.12
SU-2015 🟥 🟧	14.98
FBK 🟥	15.95
UFAL ⬛	21.89

↑

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

Evaluation Results - *EnDe*

System Ranking	**mTER HE Set 5 PErefs**
UEDIN	13.31
KIT	14.12
SU-2015	14.98
FBK	15.95
UFAL	21.89

← **- 5.94 (Δ= 27%)**

Evaluation Results - *EnDe*

System Ranking	**mTER HE Set 5 PErefs**	**HTER HE Set tgt PEref**	**TER HE Set ref**	**TER Test Set ref**
UEDIN	13.31	21.72	52.405	52.016
KIT	14.12	22.29	52.966	52.471
SU-2015	14.98	*21.09*	*51.150*	*51.130*
FBK	15.95	25.42	*51.881*	*51.561*
UFAL	21.89	28.82	57.415	57.084

Evaluation Results - *EnDe*

System Ranking	mTER HE Set 5 PErefs	HTER HE Set tgt PEref	TER HE Set ref	TER Test Set ref
UEDIN	13.31	21.72	52.405	52.016
KIT	14.12	22.29	52.966	52.471
SU-2015	14.98	*21.09*	*51.150*	*51.130*
FBK	15.95	25.42	*51.881*	*51.561*
UFAL	21.89	28.82	57.415	57.084

TER reduction

Evaluation Results - *EnDe*

System Ranking	**mTER HE Set 5 PErefs**	**HTER HE Set tgt PEref**	**TER HE Set ref**	**TER Test Set ref**
UEDIN	13.31	21.72	52.405	52.016
KIT	14.12	22.29	52.966	52.471
SU-2015	14.98	*21.09*	*51.150*	*51.130*
FBK	15.95	25.42	*51.881*	*51.561*
UFAL	21.89	28.82	57.415	57.084
Rank corr.		0.70	0.20	0.20

Spearman's Rank Coefficient

Evaluation Results - *EnFr*

System Ranking	mTER HE Set 5 PErefs
UEDIN	12.41
FBK	12.98
MMT	19.50
GT	19.98
PJAIT-2015	21.90

Evaluation Results - *EnFr*

System Ranking	**mTER HE Set 5 PErefs**
UEDIN 🟩	12.41
FBK 🟩	12.98
MMT 🟧	19.50
GT 🟧	19.98
PJAIT-2015 🟥	21.90

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

Evaluation Results - *EnFr*

System Ranking	mTER HE Set 5 PErefs
UEDIN 🟩	12.41
FBK 🟩	12.98
MMT 🟧	19.50
GT 🟧	19.98
PJAIT-2015 🟥	21.90

← - 6.52 (Δ= 33%)

Evaluation Results - *EnFr*

System Ranking	mTER HE Set 5 PErefs	HTER HE Set tgt PEref	TER HE Set ref	TER Test Set ref
UEDIN	12.41	17.89	43.456	44.457
FBK	12.98	18.51	*42.723*	*43.963*
MMT	19.50	25.18	48.151	49.456
GT	19.98	25.29	48.799	49.820
PJAIT-2015	21.90	28.28	*48.091*	*49.153*

Evaluation Results - *EnFr*

System Ranking	mTER HE Set 5 PErefs	HTER HE Set tgt PEref	TER HE Set ref	TER Test Set ref
UEDIN	12.41	17.89	43.456	44.457
FBK	12.98	18.51	*42.723*	*43.963*
MMT	19.50	25.18	48.151	49.456
GT	19.98	25.29	48.799	49.820
PJAIT-2015	21.90	28.28	*48.091*	*49.153*

TER reduction

Evaluation Results - *EnFr*

System	**mTER**	**HTER**	**TER**	**TER**
Ranking	**HE Set**	**HE Set**	**HE Set**	**Test Set**
	5 PErefs	**tgt PEref**	**ref**	**ref**
UEDIN	12.41	17.89	43.456	44.457
FBK	12.98	18.51	*42.723*	*43.963*
MMT	19.50	25.18	48.151	49.456
GT	19.98	25.29	48.799	49.820
PJAIT-2015	21.90	28.28	*48.091*	*49.153*
Rank corr.		1.00	0.60	0.60

Spearman’s Rank Coefficient

Future plans (under construction)

- Make SLT task more attractive
  - Add lectures less similar to written language
  - Lower entry barrier of task (provide ASR component)
  - Provide more training data

Future plans (under construction)

- Make SLT task more attractive
  - Add lectures less similar to written language
  - Lower entry barrier of task (provide ASR component)
  - Provide more training data

- Go where the fundings are …
  - Add Asian languages to our tasks (Japanese,...)
  - Look for new tasks

Future plans (under construction)

- Make SLT task more attractive
  - Add lectures less similar to written language
  - Lower entry barrier of task (provide ASR component)
  - Provide more training data

- Go where the fundings are …
  - Add Asian languages to our tasks (Japanese,...)
  - Look for new tasks

- Collect ideas and opinions during the workshop
  - Informal chats (please tell us what do you think)
  - Panel discussion tomorrow

Credits

- **Language resources**
  - TED LLC, USA (TED Talk data)
  - Qatar Computing Research Institute (QED Talk data)
  - Microsoft (MSLT data)
  - Conference of Machine Translation (Giga and news data)
  - DFKI, Germany (United Nations data)
- **Funding**
  - H2020 CSA CRACKER (Human evaluation)

Questions?