Overview of the IWSLT 2012 Evaluation Campaign

Marcello Federico, FBK-irst, Italy
Mauro Cettolo, FBK-irst, Italy
Luisa Bentivogli, CELCT, Italy
Michael Paul, NICT, Japan
Sebastian Stueker, KIT, Germany

IWSLT, Hong Kong, 6-7 December 2012

Outline

- IWSLT Evaluations
- TED Task
  - Specifications
  - Evaluation
  - Results
  - Conclusions
- OLYMPICS Task

  <same items but shorter>
- Conclusions

IWSLT: almost ten years old!

2004 Kyoto, Japan

2005 Pittsburgh, USA

2006 Kyoto, Japan

2007 Trento, Italy

2008 Honolulu, USA

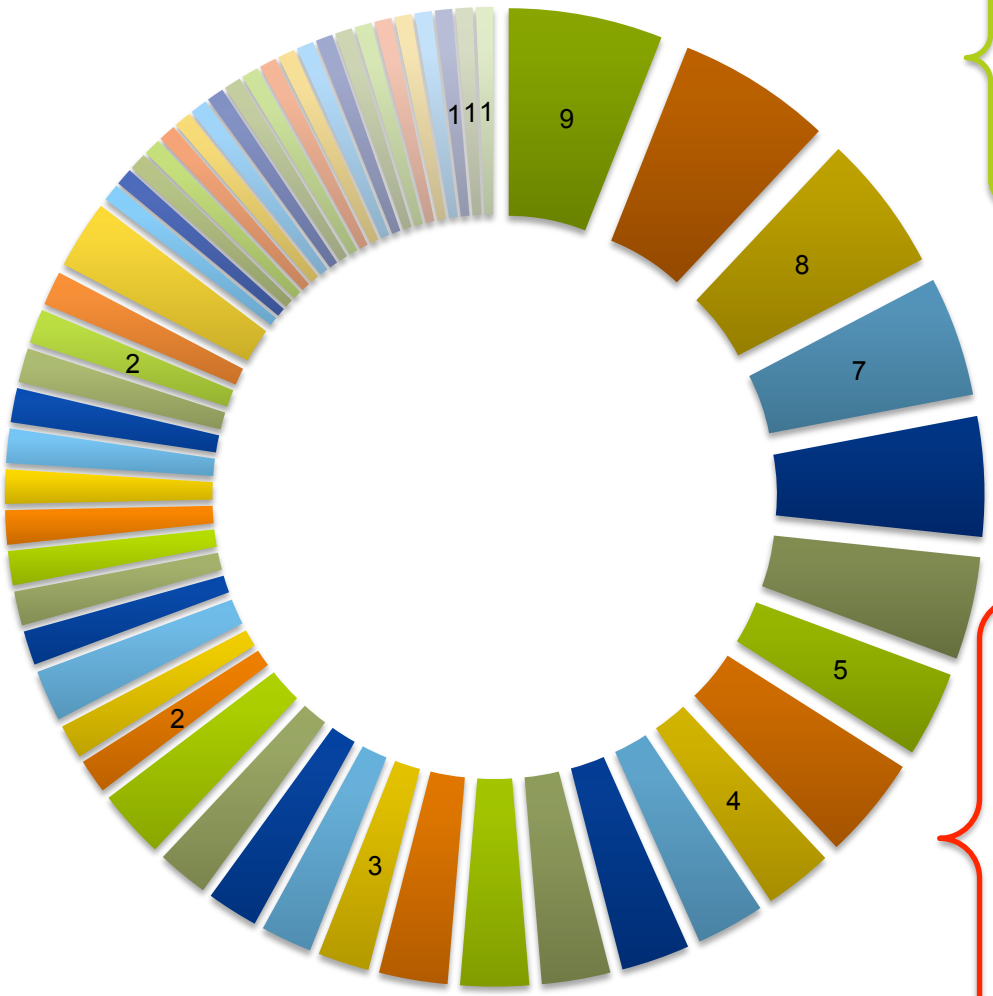
2009 Tokyo, Japan

2010 Paris, France

2011 San Francisco, USA

2012 Hong Kong, China

IWSLT Evaluation: record of participants



55 distinct participants in 9 years

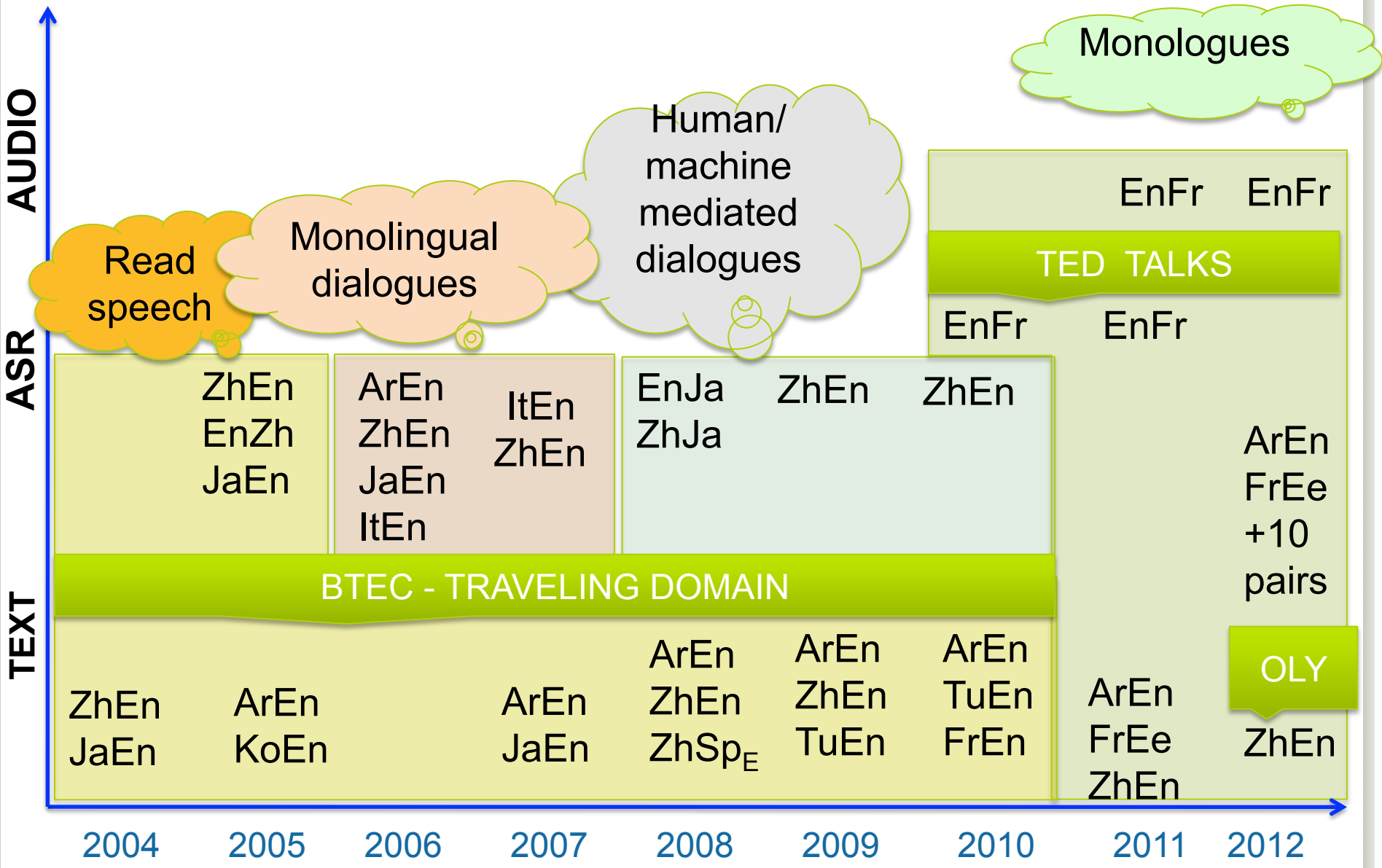
≥4X

- FBK-irst
- KIT/CMU
- MIT & Airforce
- NLPR-CAS
- IBM
- ICT-CAS
- I2R
- NICT
- RWTH
- DCU/CNGL
- LIG
- UPC- TALP
- NTT
- LIUM

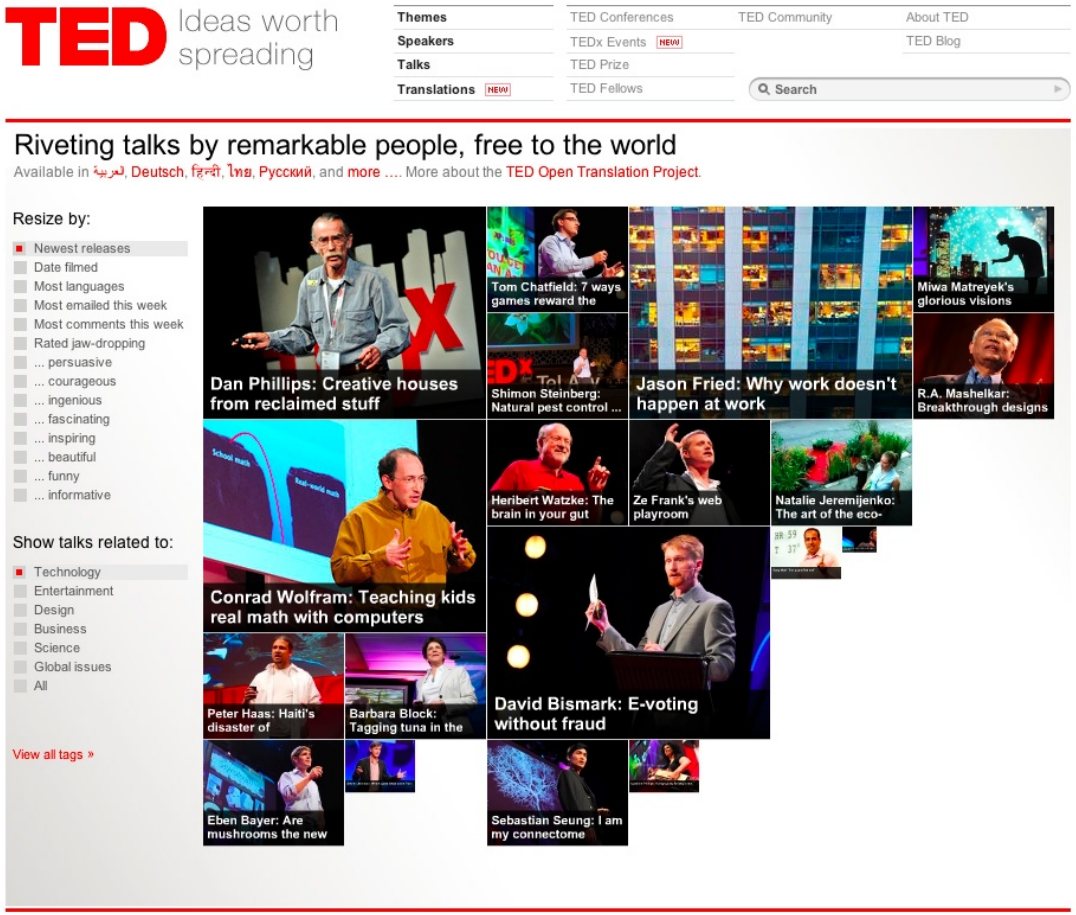
= 1x

- U. Tokyo
- U. Tottori
- TUBITAK
- CLIPS
- MSR US
- U. Xiamen
- U. Pohang
- U.P. Valencia
- HIT
- U. Kyoto
- Sehda Inc.
- OKI Ltd.
- JHU
- Sharp Labs Europe Ltd.
- U. Maryland
- Toshiba China
- U. Singapore
- POSTECH
- PJIIT
- MIRACL Lab
- DFKI
- U. Washington
- GREYC
- ISI-USC
- U. Edinburgh
- HKUST
- INESC
- Apptek, Inc
- LIMSI
- NAIST
- RALI
- NGKUT
- UNED
- AT&T
- NUDT
- U.Q.M. London
- Barcelona Media
- CMU Qatar
- RACAI
- U. CEU-Cardenal Herrera
- MSR Asia

IWSLT: tasks and languages

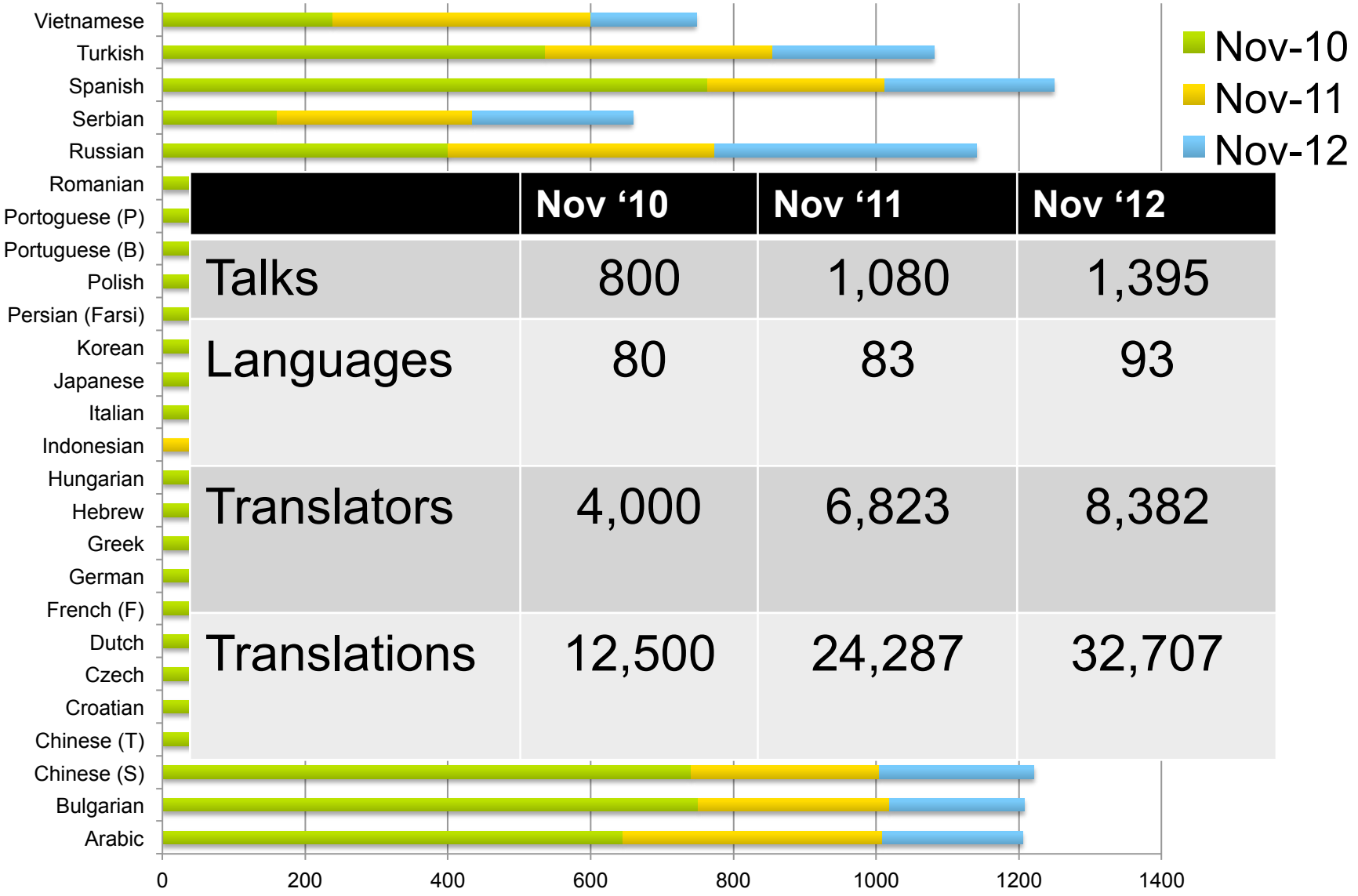


TED Talks



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

TED Talks Translations (from English)



	Nov '10	Nov '11	Nov '12
Talks	800	1,080	1,395
Languages	80	83	93
Translators	4,000	6,823	8,382
Translations	12,500	24,287	32,707

Human task: subtitling and translating

- ✓ audio is partitioned
- ✓ speech is segmented into sentences and transcribed
- ✓ translators works on the segmented transcript
- ✓ ideal translation unit is the single caption
- ✓ possible word reordering across 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 annotation 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 annotation of non-speech events
  - Subtitling and translating 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 annotation of non-speech events
  - Subtitling and translating in real-time

TED Tracks

- **Automatic Speech Recognition (ASR)**
  - transcription of talks from audio to text
  - English (E)
- **Spoken Language Translation (SLT)**
  - automatic translation of talks from audio (or ASR output) to text
  - English-French (EF)
- **Machine Translation (MT)**
  - Automatic translation of talks from text to text
  - English-French (EF), Arabic-English (AE) +

    X to English unofficial pairs

    X= Chinese, Dutch, German, Polish, Portuguese(B),

    Romanian, Russian, Slovak, Slovene, Turkish.

TED Task Resources

- **Speech**
  - any *publicly available* data recorded before 31 Dec ’10
- **Bilingual texts**
  - TED: for all language pairs
  - Multi-UN: EF(251Mw), CE (200Mw), AE (220Mw)
  - WMT: EF(800Mw)
- **Monolingual texts**
  - TED: E(2.4Mw), F(2.4Mw)
  - LDC: Giga English, Giga French
  - Google Book N-grams of English and French
- **Dev sets**
  - ASR: 8 talks (segmented and transcribed)
  - SLT: + ASR outputs (1-best and lattices)
  - MT: 8+11 talks with one reference

TED Task Specifications

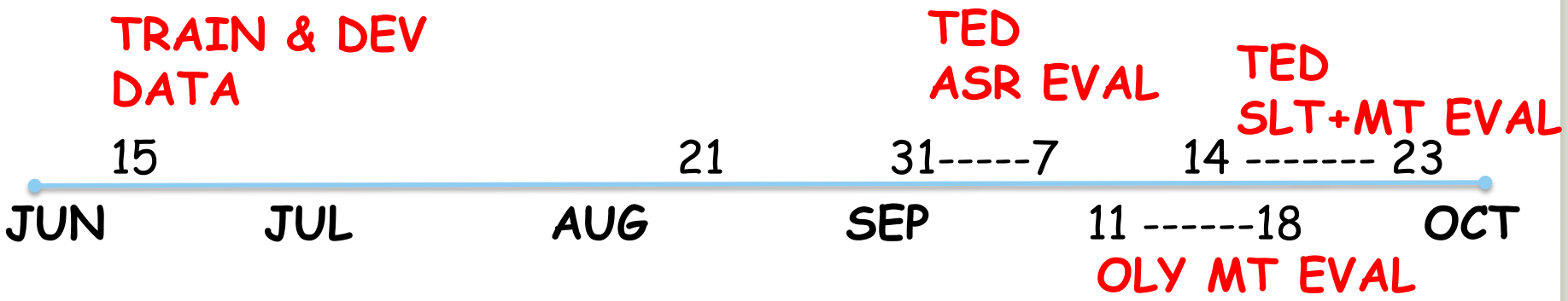
Conditions	ASR	SLT	MT
Input: Pre-segmented	yes	yes	yes
Input: Cased & Punctuated		no	yes
Output: Cased & Punctuated	no	yes	yes
Automatic evaluation (1)	yes	yes	yes
Human evaluation (official)		yes	yes

Metrics	ASR	SLT	MT
WER	✔	✔	✔
BLEU, NIST		✔	✔
METEOR, GTM		✔	✔
PER, TER		✔	✔

(1) Prepared non trivial reference baselines for all MT directions.

Evaluation Schedule

- **Timeline**



- **Submissions**
  - One primary run, multiple secondary runs
  - Runs both on TST12 and TST11 (progress test)
  - Scores computed by evaluation server

TED Task Participants

	Group	System
IT	Fondazione Bruno Kessler	FBK
DE	Karlsruhe Institute of Technology	KIT
FR	Laboratory of Informatics of Grenoble	LIG
US	Mass. Institute of Technology/Air Force Research Lab.	MIT-LL
JP	Nara Institute of Science and Technology	NAIST
JP	National Institute of Communications Technology	NICT
DJ	KIT & NAIST Collaboration	KIT-NAIST
TR	Center of Research for Advanced Technologies	TUBIKAT
PL	Polish-Japanese Institute of Information Technology	PJIIT
RO	Research Institute for AI of the Romanian Academy	RACAI
DE	Rheinisch-Westfälische Technische Hochschule Aachen	RWTH
UK	University of Edinburgh	UEDIN

TED Task Submissions

	ASR EN	SLT EF	MT EF	Ar	De	Nl	Pl	Pt	Ro	Ru	Sk	Tr	Zh
FBK	✔		✔	✔	✔	✔					✔	✔	
KIT	✔	✔	✔										
LIG			✔										
MIT	✔	✔	✔	✔									
NAIST	✔		✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	
NICT	✔									✔			✔
PJIIT							✔						
RACAI									✔				
RWTH	✔	✔	✔	✔	✔						✔		✔
TUBIKAT				✔								✔	
UEDIN	✔	✔	✔		✔								
TOTAL	7	4	7	5	4	2	2	1	2	2	3	3	2

Subjective Evaluation for SLT/MT

Judges recruited through crowd sourcing
Applied pair-wise MT output comparisons
Adopted new tournament scheme system for ranking

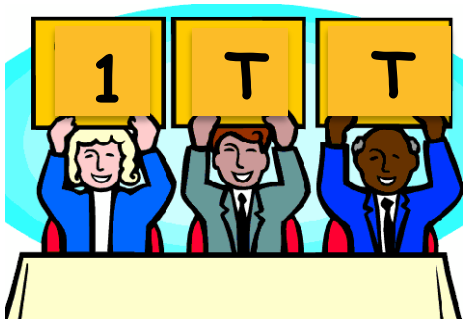
Pair-wise comparison (sentence)

Ref: Think about this as a pixel, a flying pixel.

Sys 1: This is a pixel, a pixel.
Sys 2: Take this as a pixel, a flight pixel.

and the winner is ... System 2!

Pair-wise comparison (subset of test set)



Ref: <sentence 1>
Sys 1:<output 1>
Sys 2:<output 1>

Ref: <sentence 2>
Sys 1:<output 2>
Sys 2:<output 2>

Ref: <sentence 3>
Sys 1:<output 3>
Sys 2:<output 3>

Ref: <sentence 4>
Sys 1:<output 4>
Sys 2:<output 4>

• • •

Ref: <sentence 400>
Sys 1:<output 400>
Sys 2:<output 400>

Wins: 1 = 33%, 2= 44%, T 23%

The winner is System 2!

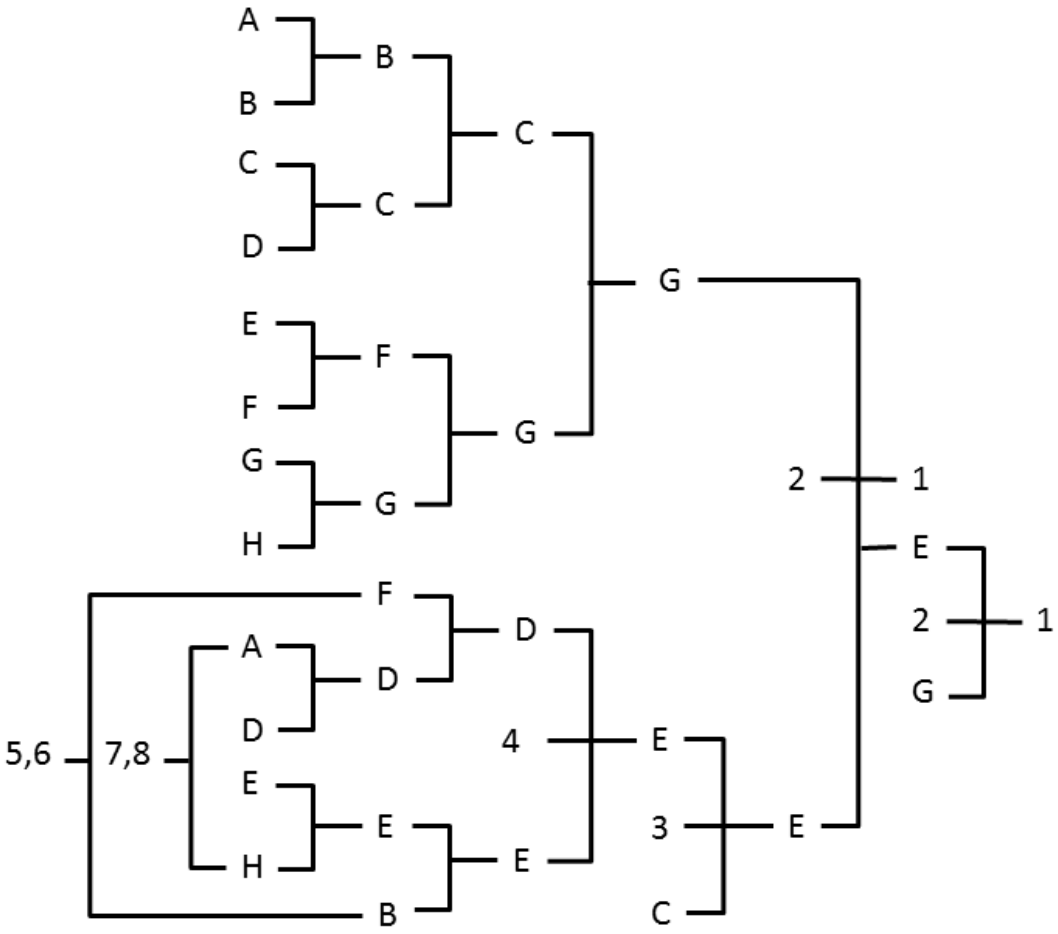
TED Task: Subjective Ranking

- Performed on the progress test set *(tst2001)*
  - to measure progress wrt 2011 systems
- New tournament scheme:
  - from ***Round Robin*** (RR) to ***Double Elimination*** (DE)
  - as reliable as RR
    - Tested on 2011 subjective evaluation data

      (SLT-EF, MT-EF,MT-AE,MT-CE tasks)
    - DE and RR rankings were the same on all tasks
  - DE scheme is more cost effective than RR
    - 16 vs. 28 pairwise comparisons (with 8 systems)

DE Tournament

8-player scheme:



- Use BLEU as ranking seed and start with:

  A-B C-D E-F G-H

  P_1-P_8 ; P_5-P_4 ; P_3-P_6 ; P_7-P_2

- Double elimination: each player must lose twice to be kicked out

- Consolation play-offs at each stage of the tournament to rank all players

DE tournament scheme

- Seed: BLEU scores on the entire 2011 test set
- Ties: resolved by counting micro “win” judgments instead of “win” labels from majority voting
- Repeated matches: use result of first match
- Results for matches involving 2011 systems were taken from IWSLT 2011 evaluation data

Subjective rankings

- **SLT-EF**: all 4 runs of 2012 + best 4 runs of 2011
- **MT-AE**: all 5 runs of 2012 + best 2 runs of 2011 + 2012 baseline
- **MT-EF**: two subsequent tournaments
  - (1) find best 4 runs among 2012 submissions
  - (2) rank top 4 runs of 2012 + top 4 runs of 2011

TASK	#SYSTEMS	#MATCHES	IAA *(Fleiss k)*
SLT-EF	8	***15*** (1 repeated match)	0.2263
MT-AE	8	***14*** (2 repeated matches)	0.2496
MT-EF	12	***22*** (6 repeated matches)	0.2861

TED Task Results: ASR

Run	TST11 WER%	TST12 WER%
NICT	**10.9**	**12.1**
KIT-NAIST	12.0	12.4
KIT	12.0	12.7
MITLL	11.1	13.3
RWTH	13.4	13.6
UEDIN	12.4	14.4
FBK	15.4	16.8

Best WER on TST11 was 13.5% = 19% improvement!

Excerpt of ASR outputs [tst2011]

```
my name is joshua  walters
i  mean if you are whoppers
*   *   *   *   *  walters

i am a performer
i am a performer
i am a performer

but as far  as  being a  performer i am also
 *   *  and partly *      performer i am also …
 *   *  * and with this fiery performer i am also …
```

errors and fixes

(IWSLT ASR TST11, Talk 1178,
ref, best ‘11, best ’12)

Excerpt of ASR outputs [tst2011]

```
i reframe that as a positive because the crazier i get
a refrain that as a positive because the crazy   i get
i refrain that as a positive because the crazier i get

when i was sixteen in  san francisco   i had my
 and i was sixteen and severance cisco *  had my
when i was sixteen in  san francisco   i had my

   you know scary but actually there's no amount of
so you know scary but actually there's no matter *
   a   no   scary but actually there's no matter *
```

(IWSLT ASR TST11, Talk 1178,
ref, best '11, best '12)

TED Task Results: SLT English-French

BLEU Ranking
(used for tournament seeding)

Ranking	System	BLEU score
1	KIT_{12}	28.86
2	LIUM_{11}	28.23
3	RWTH_{12}	27.85
4	KIT_{11}	26.78
5	RWTH_{11}	26.76
6	UEDIN_{12}	26.54
7	LIG_{11}	24.85
8	MITLL_{12}	24.27

TED Task Results: MT English-French

Find top 4 systems among 2012 participants …

BLEU Ranking
(used for tournament seeding)

Ranking	System	BLEU score
1	UEDIN$_{12}$	39.01
2	RWTH$_{12}$	38.66
3	KIT$_{12}$	38.49
4	NAIST$_{12}$	37.90
5	FBK$_{12}$	37.43
6	LIG$_{12}$	36.88
7	BASELINE$_{12}$	33.90
8	MITLL$_{12}$	31.44

TED Task Results: MT English French

… and let them compete against top 4 runs of 2011

BLEU Ranking
(used for tournament seeding)

Ranking	System	BLEU score
1	UEDIN_{12}	39.01
2	RWTH_{12}	38.66
3	KIT_{12}	38.49
4	KIT_{11}	37.65
5	LIG_{12}	36.88
6	LIMSI_{11}	36.49
7	MITLL_{11}	35.28
8	DFKI_{11}	34.39

TED Task Results: MT Arabic-English

BLEU Ranking
(used for tournament seeding)

Ranking	System	BLEU score
1	RWTH_{12}	27.28
2	RWTH_{11}	26.32
3	FBK_{12}	25.46
4	FBK_{11}	24.31
5	TUBITAK_{12}	23.85
6	NAIST_{12}	23.65
7	BASELINE_{12}	22.08
8	MITLL_{12}	17.99

MT excerpt: Arabic-English

And his brother said, "I just want to be able to talk to Tony again. I just want to be able to communicate with him and him to be able to communicate with me." And I said, "Wait a second, isn't that -- I've seen Stephen Hawking -- don't all paralyzed people have the ability to communicate via these devices?" And he said, "No, unless you're in the upper echelon, you've got really amazing insurance, you can't actually do that. These devices aren't accessible to people." (Human reference)

And he said his brother: "I just want to be Tony able to talk again. I just want to be able to communicate with him and to be able to communicate with us. And I said, "Wait a second, didn't see Stephen Hawking -- a physicist, paralyzed -- that people living in paralysis have the ability to communicate through these devices?" And he said, "No, unless you belong to the elite, you may be that I got insurance, you can't take advantage of that. These devices are not available to all people. " (Best Arabic-English MT system 2011)

And he said his brother: «Tony, I just want to be able to talk again. I just want to be able to communicate with him and to be able to communicate with us. And I said, «Wait a second, didn't see a physicist Stephen Hawking, paralyzed -- that have the ability to communicate through these devices?» And he said, «No, unless you belong to the elite, you may be that got insurance amazing you can't take advantage of that. These devices are not available to all people . " (Best Arabic-English MT system 2012)

MT excerpt: Arabic-English

And his brother said, "I just want to be able to talk to Tony again. I just want to be able to communicate with him and him to be able to communicate with me." And I said, "Wait a second, isn't that -- I've seen Stephen Hawking -- don't all paralyzed people have the ability to communicate via these devices?" And he said, "No, unless you're in the upper echelon, you've got really amazing insurance, you can't actually do that. These devices aren't accessible to people." (*Human reference*)

And he said his brother: "I just want to be Tony able to talk again. I just want to be able to communicate with him and [him] to be able to communicate with us. And I said, "Wait a second, didn't see Stephen Hawking -- a physicist, paralyzed -- that people living in paralysis have the ability to communicate through these devices?" And he said, "No, unless you belong to the elite, you may be that I got insurance, you can't take advantage of that. These devices are not available to all people." (*Best Arabic-English MT system 2011*)

And he said his brother: «Tony, I just want to be able to talk again. I just want to be able to communicate with him and [him] to be able to communicate with us. And I said, «Wait a second, didn't see a physicist Stephen Hawking, paralyzed -- that have the ability to communicate through these devices?» And he said, «No, unless you belong to the elite, you may be that got insurance$_2$ amazing$_1$ you can't take advantage of that. These devices are not available to all people ." (*Best Arabic-English MT system 2012*)

MT excerpts: German/Portuguese-English

And his brother said, "I just want to be able to talk to Tony again. I just want to be able to communicate with him and him to be able to communicate with me." And I said, "Wait a second, isn't that -- I've seen Stephen Hawking -- don't all paralyzed people have the ability to communicate via these devices?" And he said, "No, unless you're in the upper echelon, you've got really amazing insurance, you can't actually do that. These devices aren't accessible to people."

(*Human reference*)

And his brother said, "I want to talk with Tony. I just want to share with him. " And I said, "Wait a minute, isn't it - I saw Stephen Hawking - can't communicate all paralyzed people using these devices? » And he said, "No, no, it's only when you get to the higher class and has a remarkable insurance, you can do that. Normal people are not available for these devices.

(*Best German-English MT system 2012*)

The brother said, "I just want to go back to talk to Tony. I just wanted to be able to communicate with him, and he with me. " And I said, "Wait a minute, it's not that -- I saw Stephen Hawking -- people with paralysis can communicate with these devices, can't?" And he said, "No, unless you're of his top, or to have a life insurance fantastic, in fact, it's not possible. People don't have access to these devices."

(*Best Portuguese-English MT system 2012*)

MT excerpts: German/Portuguese-English

And his brother said, "I just want to be able to talk to Tony again. I just want to be able to communicate with him and him to be able to communicate with me." And I said, "Wait a second, isn't that -- I've seen Stephen Hawking -- don't all paralyzed people have the ability to communicate via these devices?" And he said, "No, unless you're in the upper echelon, you've got really amazing insurance, you can't actually do that. These devices aren't accessible to people."

(*Human reference*)

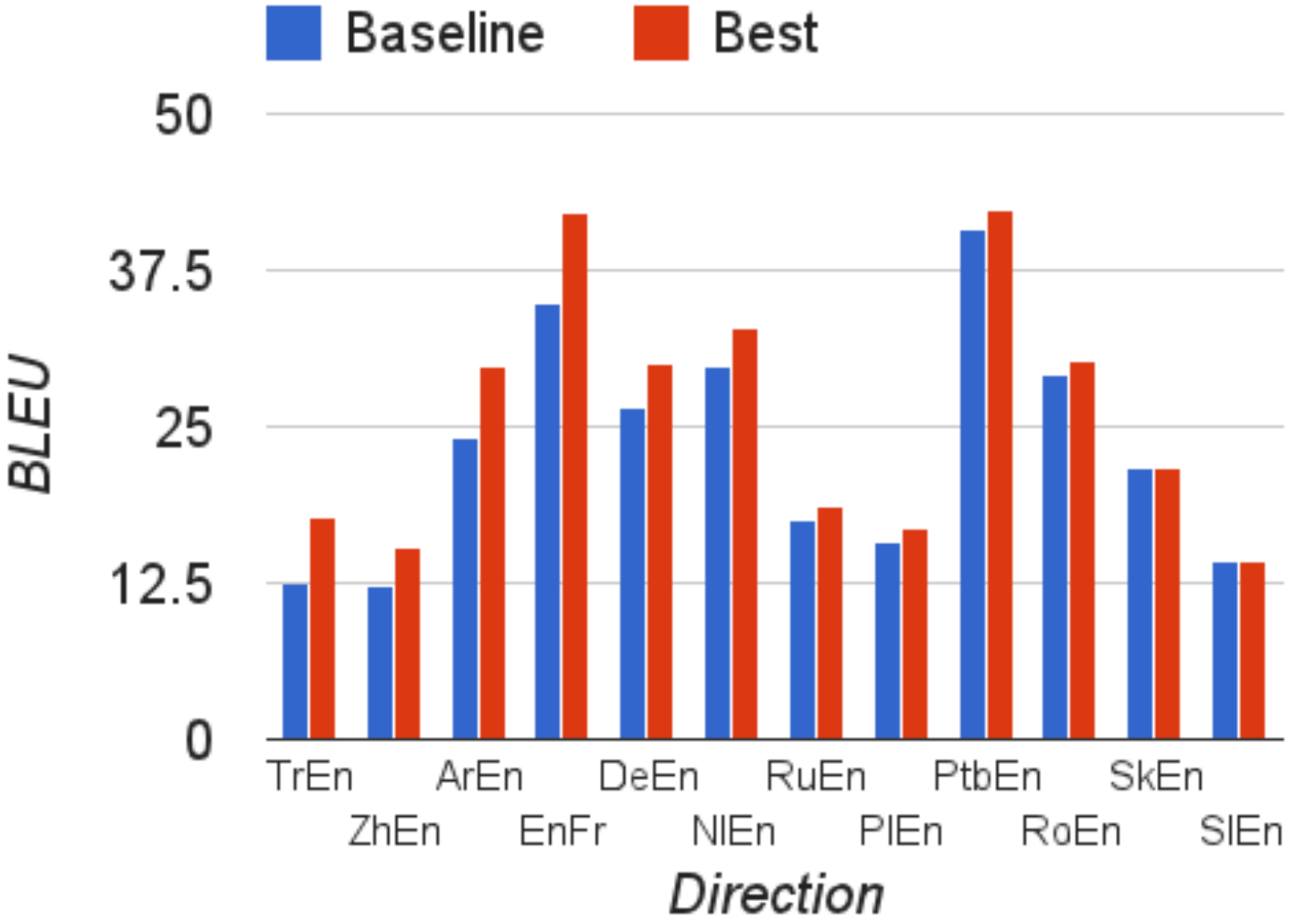
And his brother said, "I want to talk with Tony. I just want to share with him. " And I said, "Wait a minute, isn't it - I saw Stephen Hawking - can't communicate$_4$ all$_1$ paralyzed$_2$ people$_3$ using these devices? » And he said, "No, no, it's only when you get to the higher class and has a remarkable insurance, you can do that. Normal$_7$ people$_8$ are$_3$ not$_4$ available$_5$ for$_6$ these$_1$ devices$_2$.

(*Best German-English MT system 2012*)

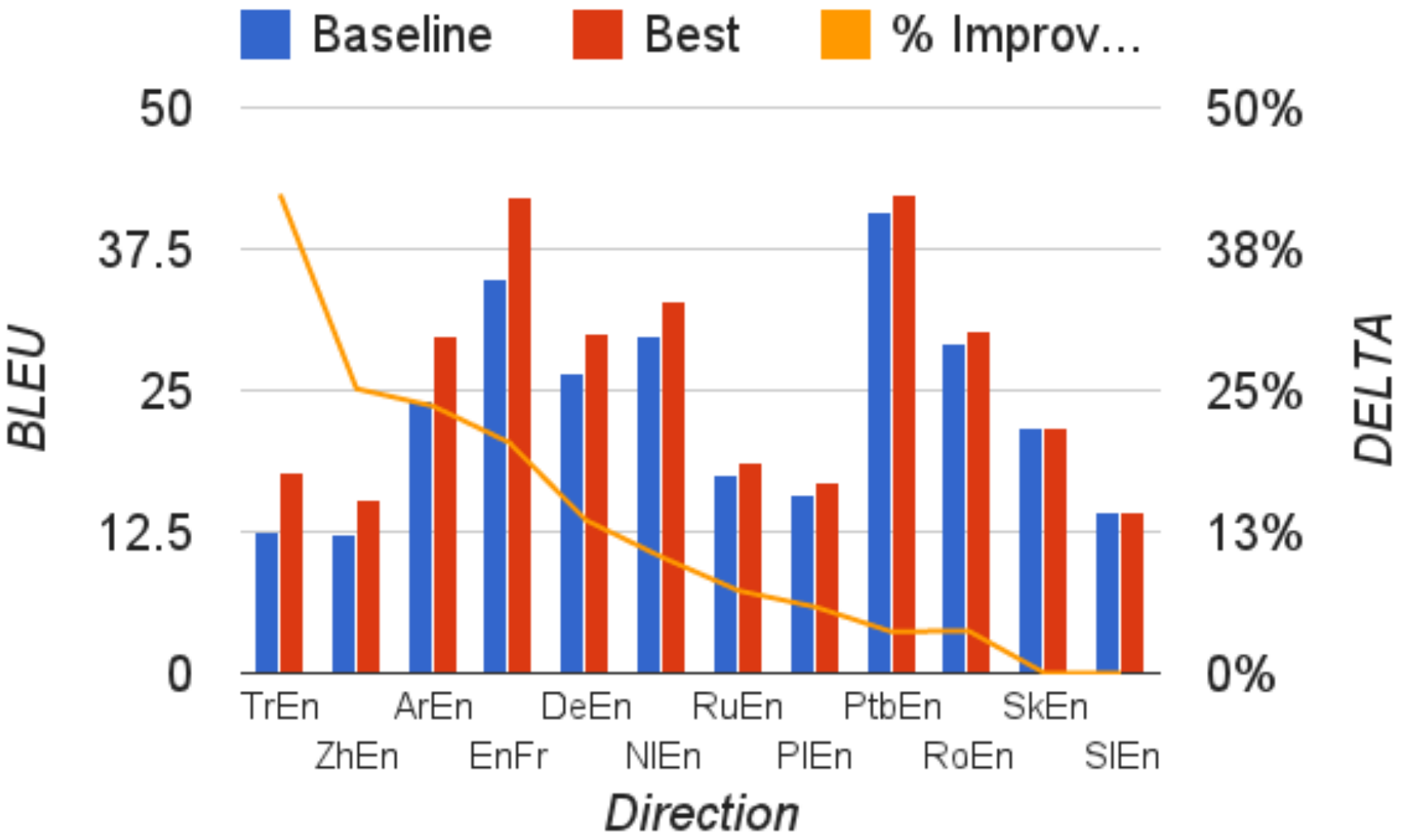
The brother said, "I just want to go back to talk to Tony. I just wanted to be able to communicate with him, and he with me. " And I said, "Wait a minute, it's not that -- I saw Stephen Hawking -- people with paralysis can communicate with these devices, can't [they]?" And he said, "No, unless you're of his top, or to have a life$_2$ insurance$_3$ fantastic$_1$, in fact, it's not possible. People don't have access to these devices."

(*Best Portuguese-English MT system 2012*)

MT Results: official and unofficial tracks



MT Results: official and unofficial tracks



TED Task Conclusions

- Three tracks: ASR (en), SLT (en-fr) and MT (en-fr,ar-en)
  - 12 teams, 44 primary runs
- Subjective evaluation with double-elimination tournament
- Evaluation with new and progress evaluation sets
  - We measures significant improvements wrt 2011
- New unofficial MT tracks with 10 language pairs
- Released TED parallel data for new languages pairs
- Prepared not-trivial baseline systems for all directions
  - Best MT systems were mostly better than baseline

OLYMPICS Task

Dialog Translation:

- human dialogs in travel situations
- translation of transcribed dialogs
- small vocabulary task
- Olympics domain
- Chinese to English

Challenges:

- sentence structure differences (non-parallel sentences)
- out-of-vocabulary words

Participants:

- 4 teams, 4 primary and 4 contrastive runs
- SMT architectures:
  + phrase-based (HIT, NICT) and syntax-based (POSTECH)
  + syntax-based EBMT (KYOTO-U)

OLYMPICS Task: Language Resources

Olympic Trilingual Corpus (HIT)

+ a multilingual corpus covering 5 domains (traveling, dining, sports, traffic and business) closely related to the Beijing 2008 Olympic Games
+ dialogs, example sentences, web articles, language teaching materials
+ train: 50k, develop: 2k, evaluation: 1k
  ⇨ single reference translations only

Basic Travel Expression Corpus (BTEC)

+ a multilingual speech corpus containing tourism-related sentence
+ training and evaluation data sets of previous IWSLT tasks
+ train: 20k, develop: 3k
  ⇨ up to 16 reference translations

OLYMPICS Task: Specifications

Evaluation Specifications:

» case-sensitive with punctuation included
» *automatic evaluation* : all primary and contrastive runs
» *human assessment*: primary runs only

Evaluation:

Objective evaluation (BLEU scores)
Subjective Ranking (Pair-wise Comparison, Round Robin)
Subjective Adequacy

+ human judgments : How much information of input is retained in translation?
(5: *all* / 4: *most* / 3: *much* / 2: *little* / 1: *none*)
+ median score of all evaluated sentences of a single system
+ with / without taking into account dialog context

OLYMPICS Task: Results

Objective

System	BLEU%
HIT	19.17
NAIST-NICT	16.95
KYOTO-U	12.79
POSTECH	12.16

Subjective ranking

System	Better%
HIT	38.08
NAISTNICT	30.25
KYOTO-U	21.50
POSTECH	8.50

Subjective adequacy

System	w/o context	with context
HIT	3.17	3.42
NAISTNICT	3.00	
KYOTO-U	2.90	
POSTECH	2.49	

OLYMPICS Task: Conclusions

- Four participants:
- Consistent rankings were obtained for human assessment
- Differences between all system are statistically significant
- Findings:
  - Improvement of sentence alignment quality via filtering
  - No improvement by adding Wikipedia data
  - Phrase-based SMT better than tree-based SMT

Thank you

Credits

- **IWSLT Evaluation Team**

  Christian Girardi, Giovanni Moretti,

  Michael Paul, Sebastian Stüker,

- **Language resources**
  - TED LLC, USA (Talk data)
  - Workshop Machine Translation (Giga and news data)
  - DFKI, Germany (United Nations data)
- **Funding**
  - EUBRIDGE IST 287658
  - Grant by European Association for Machine Translation
  - Concept for the Future, German Excellence Initiative

Questions?