

# Machine Translation Marathon 2015: MT Evaluation Lecture

Yvette Graham



# Lecture Outline

- ① Automatic Evaluation
- ② Human Evaluation
- ③ MT System Significance Tests
- ④ MT Metric Evaluation
- ⑤ MT Quality Estimation Evaluation

# Automatic Evaluation

Automatic Evaluation: commonly comparison of system output translation with human-generated reference translations.

## Automatic Evaluation

Automatic Evaluation: commonly comparison of system output translation with human-generated reference translations.


ACL 2015 Proceedings:

- BLEU [Papineni et al., 2002]: 262 mentions
- HTER [Snover et al., 2005]: 114 mentions
- METEOR [Denkowski and Lavie, 2010]: 35 mentions
- TER [Snover et al., 2005]: 15 mentions (TERp: 2 TERCpp: 1)
- NIST [Doddington, 2002]: 3 citations

# BLEU

BLEU = Bilingual Evaluation ...

## un·der·stud·y

/ˈændərˌstʌdē/ 

*noun*

1. (in the theater) a person who learns another's role in order to be able to act as a replacement at short notice.

*synonyms*: [stand-in](#), [substitute](#), [replacement](#), [reserve](#), [fill-in](#), [locum](#), [proxy](#), [backup](#), [relief](#), [standby](#), [stopgap](#); [More](#)

*verb*

1. learn (a role) or the role played by (an actor).  
"he had to understudy Prospero"

Automatic MT metrics: provide as good as possible a **stand in** for a human assessment of translation quality.

# BLEU

Let  $R$  denote a set of human reference translations and  $Z$  a set of machine translations, both produced by translating a single set of (foreign language) input sentences.

Since  $|R| = |Z|$ , we refer this quantity, the number of translations in the test set, as  $m$ .

Each automatic translation,  $Z_i$ , corresponds to exactly one human reference translation,  $R_j$ .

A translation consists of a sequence of words  $\langle w_1 \dots w_n \rangle$ , and a  $k$ -gram, is a contiguous  $k$ -length sequence of words within a translation,  $\langle w_{i-(k-1)} \dots w_i \rangle$ , where  $k \leq i \leq n$ .

BLEU is computed as follows:

$$S_Z = BP_Z \cdot (P_{Z1} \dots P_{Z4})^{1/4}, \quad (1)$$

where  $BP_Z$  denotes the “brevity penalty” for machine  $Z$ .

## BLEU's Brevity Penalty

The brevity penalty,  $BP_Z$  is computed as follows:

$$BP_Z = \begin{cases} 1 & \text{if } len(Z) \geq len(R) \\ e^{(1-len(R)/len(Z))} & \text{if } len(Z) < len(R) \end{cases}$$

where  $len(Z) = \sum_{i=1}^m len(Z_i)$ , and  $len(R) = \sum_{i=1}^m len(R_i)$ , and function  $len$  returns the number of words in a given translation.

BLEU's Brevity Penalty penalises a system for producing translations that are too short in much the same way recall does in an F-score.

## BLEU's Precision Scores

The precision score,  $P_{Zk}$  is computed as follows:

$$P_{Zk} = \frac{\sum_{i=1}^m M_{Zik}}{\sum_{i=1}^m T_{Zik}}. \quad (2)$$

For the  $i$ th of perhaps  $m = 3,000$  automatic translations, we compute the  $k$ th “match count”  $M_{Zik}$  and  $k$ th “test count”,  $T_{Zik}$ , corresponding to machine  $Z$ , with  $1 \leq k \leq 4$ .

“Match count”,  $M_{Zik}$ , is the number of  $k$ -grams in translation,  $Z_i$ , that match those in the reference translation,  $R_i$ , and “test count”,  $T_{Zik}$ , is the number of  $k$ -grams in translation,  $Z_i$ .



## BLEU Example I

Given two MT systems, the first of which produces the set of translations,  $X$ , and the second the set of translations,  $Y$ , both of which are evaluated against the same set of human reference translations,  $R$ . For the purpose of illustration, let the number of translations in the test set,  $m$ , which is commonly 3,000, equal 5, and  $m = |X| = |Y| = |R| = 5$ . Translations in  $X$ ,  $Y$  and  $R$  are as follows:

$i$	Translation $X_i$	Translation $Y_i$	Human Reference $R_i$
1	The speedy red fox jumped	A quick fox jumped	The quick brown fox jumped
2	Many fingers make light work	Lots of hands make light manual labor	Many hands make light work
3	Negligible time to dislocate	No time to lose	No time to lose
4	No location like house	No place like home	No place like home
5	We are not in Missouri anymore	They are not in Kansas anymore	We are not in Kansas anymore

$P_{Zk}$  is computed for systems  $X$  and  $Y$ , for  $k = 1, 2, 3$  and 4.

## BLEU Example II

**Precision score  $P_{Z1}$ :** When computing the ( $k = 1$ ) precision score,  $P_{Z1}$ , we only consider  $k$ -grams of length 1, and compute a match count,  $M_{Z1}$ , and a test count,  $T_{Z1}$ . Matching 1-grams in translations of  $X$  and  $Y$  are underlined.

$i$	Translation $X_i$	Translation $Y_i$	Human Reference $R_i$
1	<u>The</u> speedy red <u>fox</u> <u>jumped</u> $M_{X,1,1} = 3$ and $T_{X,1,1} = 5$	A <u>quick</u> <u>fox</u> <u>jumped</u> $M_{Y,1,1} = 3$ and $T_{Y,1,1} = 4$	The quick brown fox jumped
2	<u>Many</u> fingers <u>make</u> <u>light</u> <u>work</u> $M_{X,2,1} = 4$ and $T_{X,2,1} = 5$	Lots of <u>hands</u> <u>make</u> <u>light</u> manual labor $M_{Y,2,1} = 3$ and $T_{Y,2,1} = 7$	Many hands make light work
3	Negligible <u>time</u> <u>to</u> dislocate $M_{X,3,1} = 2$ and $T_{X,3,1} = 4$	<u>No</u> <u>time</u> <u>to</u> <u>lose</u> $M_{Y,3,1} = 4$ and $T_{Y,3,1} = 4$	No time to lose
4	<u>No</u> location <u>like</u> house $M_{X,4,1} = 2$ and $T_{X,4,1} = 4$	<u>No</u> <u>place</u> <u>like</u> <u>home</u> $M_{Y,4,1} = 4$ and $T_{Y,4,1} = 4$	No place like home
5	<u>We</u> <u>are</u> <u>not</u> <u>in</u> Missouri <u>anymore</u> $M_{X,5,1} = 5$ and $T_{X,5,1} = 6$	They <u>are</u> <u>not</u> <u>in</u> <u>Kansas</u> <u>anymore</u> $M_{Y,5,1} = 5$ and $T_{Y,5,1} = 6$	We are not in Kansas anymore

## BLEU Example III

$P_{X1}$	$P_{Y1}$
$\frac{M_{X,1,1}+M_{X,2,1}+M_{X,3,1}+M_{X,4,1}+M_{X,5,1}}{T_{X,1,1}+T_{X,2,1}+T_{X,3,1}+T_{X,4,1}+T_{X,5,1}}$	$\frac{M_{Y,1,1}+M_{Y,2,1}+M_{Y,3,1}+M_{Y,4,1}+M_{Y,5,1}}{T_{Y,1,1}+T_{Y,2,1}+T_{Y,3,1}+T_{Y,4,1}+T_{Y,5,1}}$
$\frac{3+4+2+2+5}{5+5+4+4+6}$	$\frac{3+3+4+4+5}{4+7+4+4+6}$
$\frac{16}{24}$	$\frac{19}{25}$

## BLEU Example IV

**Precision score  $P_{Z_2}$  ( $k = 2$ ):** we only consider  $k$ -grams of length 2, and compute a match count,  $M_{Z_2}$ , and a test count,  $T_{Z_2}$ . Matching 2-grams in translations of  $X$  and  $Y$  are underlined:

$i$	Translation $X_i$	Translation $Y_i$	Human Reference $R_i$
1	The speedy red <u>fox jumped</u> $M_{X,1,2} = 1$ and $T_{X,1,2} = 4$	A quick <u>fox jumped</u> $M_{Y,1,2} = 1$ and $T_{Y,1,2} = 3$	The quick brown fox jumped
2	Many fingers <u>make light work</u> $M_{X,2,2} = 2$ and $T_{X,2,2} = 4$	Lots of <u>hands make</u> light manual labor $M_{Y,2,2} = 2$ and $T_{Y,2,2} = 6$	Many hands make light work
3	Negligible <u>time to</u> dislocate $M_{X,3,2} = 1$ and $T_{X,3,2} = 3$	<u>No time</u> to lose $M_{Y,3,2} = 3$ and $T_{Y,3,2} = 3$	No time to lose
4	No location like house $M_{X,4,2} = 0$ and $T_{X,4,2} = 3$	<u>No place</u> like home $M_{Y,4,2} = 3$ and $T_{Y,4,2} = 3$	No place like home

# BLEU Example V

$i$	Translation $X_i$	Translation $Y_i$	Human Reference $R_i$
5	<p><u>We are not</u> in Missouri anymore            _____            _____</p> <p><math>M_{X,5,2} = 3</math> and <math>T_{X,5,2} = 5</math></p>	<p>They <u>are not</u> in Kansas anymore            _____            _____            _____</p> <p><math>M_{Y,5,2} = 4</math> and <math>T_{Y,5,2} = 5</math></p>	<p>We are not in Kansas anymore</p>

## BLEU Example VI

$P_{X2}$	$P_{Y2}$
$\frac{M_{X,1,2}+M_{X,2,2}+M_{X,3,2}+M_{X,4,2}+M_{X,5,2}}{T_{X,1,2}+T_{X,2,2}+T_{X,3,2}+T_{X,4,2}+T_{X,5,2}}$	$\frac{M_{Y,1,2}+M_{Y,2,2}+M_{Y,3,2}+M_{Y,4,2}+M_{Y,5,2}}{T_{Y,1,2}+T_{Y,2,2}+T_{Y,3,2}+T_{Y,4,2}+T_{Y,5,2}}$
$\frac{1+2+1+0+3}{4+4+3+3+5}$	$\frac{1+2+3+3+4}{3+6+3+3+5}$
$\frac{7}{19}$	$\frac{13}{18}$

## BLEU Example VII

**Precision score  $P_{Z3}$  ( $k = 3$ ):** we only consider  $k$ -grams of length 3, and compute a match count,  $M_{Z3}$ , and a test count,  $T_{Z3}$ . Matching 3-grams in translations of  $X$  and  $Y$  are underlined:

$i$	Translation $X_i$	Translation $Y_i$	Human Reference $R_i$
1	The speedy red fox jumped $M_{X,1,3} = 0$ and $T_{X,1,3} = 3$	A quick fox jumped $M_{Y,1,3} = 0$ and $T_{Y,1,3} = 2$	The quick brown fox jumped
2	Many fingers <u>make light work</u> $M_{X,2,3} = 1$ and $T_{X,2,3} = 3$	Lots of <u>hands make light</u> manual labor $M_{Y,2,3} = 1$ and $T_{Y,2,3} = 5$	Many hands make light work
3	Negligible time to dislocate $M_{X,3,3} = 0$ and $T_{X,3,3} = 2$	<u>No time to lose</u> $M_{Y,3,3} = 2$ and $T_{Y,3,3} = 2$	No time to lose
4	No location like house $M_{X,4,3} = 0$ and $T_{X,4,3} = 2$	<u>No place like home</u> $M_{Y,4,3} = 2$ and $T_{Y,4,3} = 2$	No place like home

## BLEU Example VIII

$i$	Translation $X_i$	Translation $Y_i$	Human Reference $R_i$
5	<p><u>We are not</u> in Missouri anymore _____</p> <p><math>M_{X,5,3} = 2</math> and <math>T_{X,5,3} = 4</math></p>	<p>They <u>are not in</u> Kansas anymore _____</p> <p><math>M_{Y,5,3} = 3</math> and <math>T_{Y,5,3} = 4</math></p>	<p>We are not in Kansas anymore</p>



# BLEU Example IX

$P_{X3}$	$P_{Y3}$
$\frac{M_{X,1,1}+M_{X,2,1}+M_{X,3,1}+M_{X,4,1}+M_{X,5,1}}{T_{X,1,1}+T_{X,2,1}+T_{X,3,1}+T_{X,4,1}+T_{X,5,1}}$ $\frac{0+1+0+0+2}{3+3+2+2+4}$ $\frac{3}{14}$	$\frac{M_{Y,1,1}+M_{Y,2,1}+M_{Y,3,1}+M_{Y,4,1}+M_{Y,5,1}}{T_{Y,1,1}+T_{Y,2,1}+T_{Y,3,1}+T_{Y,4,1}+T_{Y,5,1}}$ $\frac{0+1+2+2+3}{2+5+2+2+4}$ $\frac{8}{15}$

## BLEU Example X

**Precision score  $P_{Z_4}$  ( $k = 4$ ):** we only consider  $k$ -grams of length 4, and compute a match count,  $M_{Z_4}$ , and a test count,  $T_{Z_4}$ . Matching 4-grams in translations of  $X$  and  $Y$  are underlined:

$i$	Translation $X_i$	Translation $Y_i$	Human Reference $R_i$
1	The speedy red fox jumped $M_{X,1,4} = 0$ and $T_{X,1,4} = 2$	A quick fox jumped $M_{Y,1,4} = 0$ and $T_{Y,1,4} = 1$	The quick brown fox jumped
2	Many fingers make light work $M_{X,2,4} = 0$ and $T_{X,2,4} = 2$	Lots of hands make light manual labor $M_{Y,2,4} = 0$ and $T_{Y,2,4} = 4$	Many hands make light work
3	Negligible time to dislocate $M_{X,3,4} = 0$ and $T_{X,3,4} = 1$	<u>No time to lose</u> $M_{Y,3,4} = 1$ and $T_{Y,3,4} = 1$	No time to lose
4	No location like house $M_{X,4,4} = 0$ and $T_{X,4,4} = 1$	<u>No place like home</u> $M_{Y,4,4} = 1$ and $T_{Y,4,4} = 1$	No place like home

# BLEU Example XI

$i$	Translation $X_i$	Translation $Y_i$	Human Reference $R_i$
5	<p><u>We are not in</u> Missouri anymore</p> <p><math>M_{X,5,4} = 1</math> and <math>T_{X,5,4} = 3</math></p>	<p>They <u>are not in Kansas</u> anymore</p> <p><math>M_{Y,5,4} = 2</math> and <math>T_{Y,5,4} = 3</math></p>	We are not in Kansas anymore

## BLEU Example XII

$P_{X4}$	$P_{Y4}$
$\frac{M_{X,1,1}+M_{X,2,1}+M_{X,3,1}+M_{X,4,1}+M_{X,5,1}}{T_{X,1,1}+T_{X,2,1}+T_{X,3,1}+T_{X,4,1}+T_{X,5,1}}$ $\frac{0+0+0+0+1}{2+2+1+1+3}$ $\frac{1}{9}$	$\frac{M_{Y,1,1}+M_{Y,2,1}+M_{Y,3,1}+M_{Y,4,1}+M_{Y,5,1}}{T_{Y,1,1}+T_{Y,2,1}+T_{Y,3,1}+T_{Y,4,1}+T_{Y,5,1}}$ $\frac{0+0+1+1+2}{1+4+1+1+3}$ $\frac{4}{10}$

## BLEU Example XIII

BLEU score  $S_Z$ :

The BLEU scores for  $X$  and  $Y$  are computed as follows:

$S_X$		$S_Y$	
$BP_X \cdot$	$P_X$	$BP_Y \cdot$	$P_Y$
1 ·	$(\frac{16}{24} \cdot \frac{7}{19} \cdot \frac{3}{14} \cdot \frac{1}{9})^{1/4}$	1 ·	$(\frac{19}{25} \cdot \frac{13}{18} \cdot \frac{8}{15} \cdot \frac{4}{10})^{1/4}$
	0.2765		0.5850

## Downsides of BLEU

BLEU has been criticised for:

- Being biased in favor of systems that happen to produce translations that are superficially very similar to reference translations
- No weighting on important words, eg. the word *not* being present or absent from a translation
- Most worrying of all: there is no guarantee that an increase in BLEU score is an indicator of improved translation quality
- See Callison-Burch et al. [2006] for further details ...

# Human-targeted Automatic Metrics

Human-targeted Metric:

- Does not employ one (or more) generic reference translations for evaluation
- Instead a customized reference translation is created manually by a human assessor – one per system output translation
- Customized reference translation created with *minimal* editing by the human assessor
- Then an automatic score is computed for translations by comparison with the customized human post-edited translation as a reference

Human-targeted metrics overcome a main bias of BLEU and other metrics by comparison, eg. Human-targeted Translation Error Rate (HTER [Snover et al., 2005]), cost as they are no longer fully automatic.

# Lecture Outline

- ① Automatic Evaluation
- ② Human Evaluation
- ③ MT System Significance Tests
- ④ MT Metric Evaluation
- ⑤ MT Quality Estimation Evaluation



# Human Evaluation

Human evaluation of MT is important for evaluating:

- MT systems (since automatic metrics are imperfect)
- MT Metrics – to see investigate how imperfect the metrics are
- MT quality estimation systems – as a gold standard to compare predictions with

# Human Evaluation

WMT [Bojar et al., 2014]:

- Relative Preference Judgments – given a set of 5 translations rank them from best to worst
- Ranking of 5 systems produces 10 pairwise comparison labels
- Several proposed methods of combining pairwise comparisons to get a single score for a system and a ranking of competing systems

Problems:

- Low levels of intra- and inter-annotator agreement
- Quality controlling crowd-sourcing judgments is difficult since even expert assessors do not agree with themselves, let alone other human assessors
- Relative preference judgments not very good for evaluating segment-level metrics and quality estimation systems – discrete rel. pre. vs continuous metric and QE system scores

# Continuous Monolingual Human Annotation I

Read the text below and rate it by how much you agree that:

**The black text adequately expresses the meaning of the gray text.**

On Facebook, it's impossible to know how much of a user's profile is true.

With Facebook, it's difficult to know how many of a user profile information is true.

strongly  
disagree



strongly  
agree

[Graham et al., 2015a]

- Mean scores for systems or individual translations quickly increase in accuracy, as higher numbers of judgments are collected
- Don't suffer from information-loss the way rel.pref. judgments do
  - Rel. Pref. judgments: with the same judgment, both translations could be high quality or both could be terrible, we lose this info.

# Continuous Monolingual Human Annotation II

- By approximately 15 repeat assessments, highly accurate segment-level scores – very useful for evaluation segment-level metrics
- By around 1500 translation assessments per system, high levels of conclusivity in system rankings
- Facilitates accurate quality control of crowd-sourcing at a low cost (around \$40 per system on Mechanical Turk)
- Facilitates longitudinal evaluation (how much systems improve compared to last year)

Crowd-sourcing Human MT Assessments for **MT System Evaluations**:

<https://github.com/ygraham/crowd-alone>  
[Graham et al., 2015b]

Crowd-sourcing Human MT Assessments for **Segment-level Metrics Eval.**:

<https://github.com/ygraham/segment-mteval>  
[Graham et al., 2015a]

## Example Longitudinal Evaluation [Graham et al., 2014a] I

		CURR <sub>07</sub>	CURR <sub>12</sub>	$\Delta$ (CURR <sub>07</sub> - CURR <sub>12</sub> )	BEST <sub>07</sub>	BEST <sub>12</sub>	5-Year Gain (BEST <sub>12</sub> - BEST <sub>07</sub> + $\Delta$ )	
DE-EN	fluency	score	65.3****	57.9	7.4	52.8	55.0* (+2.2)	9.6
		n	2,164	3,381		2,242	3,253	
	adequacy	score	63.8****	52.8	11.0	46.5	49.8** (+3.3)	14.3
		n	1,458	2,175		1,454	2,193	
metrics	BLEU	38.3	26.5	11.8	21.1	23.8 (+2.7)	14.5	
	METEOR	40.3	32.7	7.6	33.4	31.7 (-1.7)	5.9	
FR-EN	fluency	score	65.9****	58.0	7.9	57.8	60.2** (+2.4)	10.3
		n	2,172	3,267		2,203	3,238	
	adequacy	score	61.0****	52.3	8.7	52.7	51.5 (-1.2)	7.5
		n	1,754	2,651		1,763	2,712	
metrics	BLEU	39.4	32.0	7.4	28.6	31.5 (+2.9)	10.3	
	METEOR	39.8	34.6	5.2	35.9	34.3 (-1.6)	3.6	
ES-EN	fluency	score	68.4****	59.2	9.2	56.7	56.7 (+0.0)	9.2
		n	1,514	2,234		1,462	2,230	
	adequacy	score	68.0****	56.9	11.1	59.0****	55.7 (-3.3)	7.8
		n	1,495	2,193		1,492	2,180	
metrics	BLEU	51.2	38.3	12.9	35.1	33.5 (-1.6)	11.3	
	METEOR	45.4	37.0	8.4	39.9	36.0 (-3.9)	4.5	
CS-EN	fluency	score	62.3****	49.9	12.4	40.8	50.5**** (+9.7)	22.1
		n	1,873	2,816		1,923	2,828	
	adequacy	score	62.4****	47.5	14.9	41.7	47.4**** (+5.7)	20.6
		n	1,218	1,830		1,257	1,855	
metrics	BLEU	52.3	25.0	27.3	25.1	22.4 (-2.7)	24.6	
	METEOR	44.7	31.6	13.1	34.3	30.8 (-3.5)	9.6	

## Example Longitudinal Evaluation [Graham et al., 2014a] II

		CURR <sub>07</sub>	CURR <sub>12</sub>	$\Delta$ (CURR <sub>07</sub> - CURR <sub>12</sub> )	BEST <sub>07</sub>	BEST <sub>12</sub>	5-Year Gain (BEST <sub>12</sub> - BEST <sub>07</sub> + $\Delta$ )	
EN-ES	fluency	score	77.2***	73.4	3.8	63.3	71.9*** (+8.6)	12.4
		<i>n</i>	2,286	3,318		2,336	3,420	
	adequacy	score	75.2***	68.1	7.1	62.5	67.2 (+4.7)	11.8
		<i>n</i>	1,410	2,039		1,399	2,112	
metrics	BLEU	48.2	38.7	9.5	29.1	35.3 (+6.2)	15.7	
	METEOR	69.9	59.6	10.3	57.0	58.1 (+1.1)	11.4	
EN-FR	fluency	score	57.1	55.2	1.9	49.5	56.4 (+6.9)	8.8
		<i>n</i>	1,008	1,645		1,039	1,588	
	adequacy	score	64.2*	61.9	2.3	57.2	62.3 (+5.1)	7.4
		<i>n</i>	1,234	1,877		1,274	1,775	
metrics	BLEU	37.2	30.8	6.4	25.3	29.9 (+4.6)	11.0	
	METEOR	59.4	52.9	6.5	50.4	52.0 (+1.6)	8.1	
EN-DE	fluency	score	52.3	54.1*	-1.8	53.7	55.5 (+1.8)	0.0
		<i>n</i>	1,317	1,993		1,308	2,022	
	adequacy	score	60.3**	57.4	2.9	58.3	58.3 (+0.0)	2.9
		<i>n</i>	1,453	2,105		1,410	2,152	
metrics	BLEU	23.6	18.7	4.9	14.6	17.2 (+2.6)	7.5	
	METEOR	44.7	39.1	5.6	36.7	38.0 (+1.3)	6.9	

## Example Longitudinal Evaluation [Graham et al., 2014a] III

		CURR <sub>07</sub>	CURR <sub>12</sub>	$\Delta$ (CURR <sub>07</sub> - CURR <sub>12</sub> )	BEST <sub>07</sub>	BEST <sub>12</sub>	5-Year Gain (BEST <sub>12</sub> - BEST <sub>07</sub> + $\Delta$ )
fluency	score	64.1	58.2	5.9	53.5	58.0 (+4.5)	10.4
	$z$	0.18	0.00	0.18	-0.16	0.00 (+0.16)	0.34
	$n$	12,334	18,654		12,513	18,579	
adequacy	score	65.0	56.7	8.3	54.0	56.0 (+2.0)	10.3
	$z$	0.18	-0.07	0.25	-0.16	-0.09 (+0.07)	0.32
	$n$	10,022	14,870		10,049	14,979	
metrics	BLEU	41.5	30.0	11.4	25.6	27.7 (+2.1)	13.5
	METEOR	49.2	41.1	8.1	41.1	40.1 (-1.0)	7.1

- On average a 10% improvement was made to MT of European language pairs over the five years since 2007–2012
- Czech to English translation by far making the greatest gain in five years

# Lecture Outline

- ① Automatic Evaluation
- ② Human Evaluation
- ③ MT System Significance Tests**
- ④ MT Metric Evaluation
- ⑤ MT Quality Estimation Evaluation



## MT System Significance Tests

Comparing BLEU scores of System A and System B, we do not want to conclude an increase in performance if such a difference in scores is likely to have occurred simply by chance!

**Issue:** automatic MT metrics such as BLEU are calculated at the *document-level*, over the totality of translations, and return a single aggregated score, not segment-level scores

**Solution:** *randomised significance testing*, over scores for sub-samples of translations

## MT System Significance Testing

Three common randomized tests for significance testing differences in MT metric scores:

- Paired bootstrap resampling (662 GS cit.) [Koehn, 2004]

## MT System Significance Testing

Three common randomized tests for significance testing differences in MT metric scores:

- Paired bootstrap resampling (662 GS cit.) [Koehn, 2004]
- Approximate randomization (99 GS cit.) [Riezler and Maxwell, 2005]

## MT System Significance Testing

Three common randomized tests for significance testing differences in MT metric scores:

- Paired bootstrap resampling (662 GS cit.) [Koehn, 2004]
- Approximate randomization (99 GS cit.) [Riezler and Maxwell, 2005]
- Bootstrap Resampling

## MT System Significance Testing

Three common randomized tests for significance testing differences in MT metric scores:

- Paired bootstrap resampling (662 GS cit.) [Koehn, 2004]
- Approximate randomization (99 GS cit.) [Riezler and Maxwell, 2005]
- Bootstrap Resampling

Possible criticism of bootstrap resampling that  $S_{H_0}$  has the same shape but a different mean than  $S_{boot}$

But it has *\*not\** been shown empirically that any of these tests has superior accuracy than any other (... see [Graham et al., 2014b] WMT paper for further details)

## Paired Bootstrap Resampling

---

Set  $c = 0$

For bootstrap samples  $b = 1, \dots, B$

If  $S_{X_b} < S_{Y_b}$

$c = c + 1$

If  $c/B \leq \alpha$

Reject the null hypothesis

---

# Bootstrap Resampling

---

Set  $c = 0$

Compute actual statistic of score differences  $S_X - S_Y$  on test data

Calculate sample mean  $\tau_B = \frac{1}{B} \sum_{b=1}^B S_{X_b} - S_{Y_b}$  over bootstrap samples  $b = 1, \dots, B$

For bootstrap samples  $b = 1, \dots, B$

Sample with replacement from variable tuples test sentences for systems  $X$  and  $Y$

Compute pseudo-statistic  $S_{X_b} - S_{Y_b}$  on bootstrap data

If  $S_{X_b} - S_{Y_b} - \tau_B \geq S_X - S_Y$

$c = c + 1$

If  $c/B \leq \alpha$

Reject the null hypothesis

---

# Approximate Randomization

---

Set  $c = 0$

Compute actual statistic of score differences  $S_X - S_Y$   
on test data

For random shuffles  $r = 1, \dots, R$

For sentences in test set

Shuffle variable tuples between systems  $X$  and  $Y$   
with probability 0.5

Compute pseudo-statistic  $S_{X_r} - S_{Y_r}$  on shuffled data

If  $S_{X_r} - S_{Y_r} \geq S_X - S_Y$

$c = c + 1$

If  $c/R \leq \alpha$

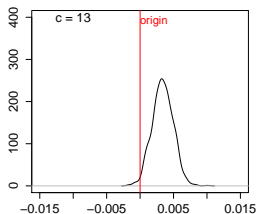
Reject the null hypothesis

---

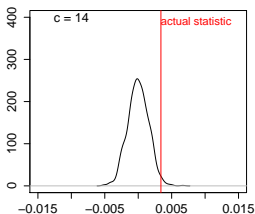


# Example Pseudo-statistic Distributions

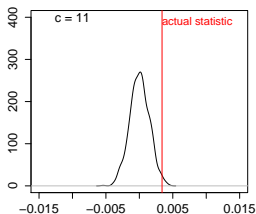
Paired Bootstrap Res. BLEU



Bootstrap Resampling BLEU



Approximate Randomization BLEU



## Evaluation Dataset

- Use translations from all participating WMT12 ES–EN and EN–ES systems (12 and 11 systems, resp.)
- Use AMT to manually annotate each translation for fluency and adequacy based on a continuous (Likert) scale, with strict annotator-level quality controls [Graham et al., 2013]
- Standardize the scores from a given annotator according to mean and standard deviation
- Final dataset: average of 1,483 (1,280) adequacy and 1,534 (1,013) fluency assessments per ES–EN (EN–ES) system

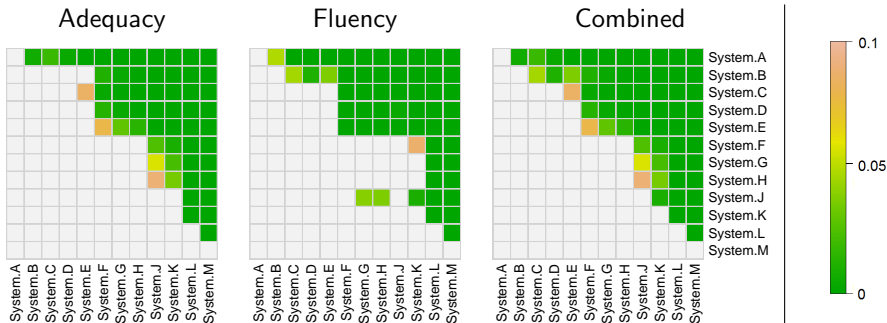
# Evaluation Methodology

- Evaluate each pair of systems separately for:
  - ① adequacy
  - ② fluency
  - ③ combined adequacy–fluency (if no significant difference in adequacy, use fluency as fallback)

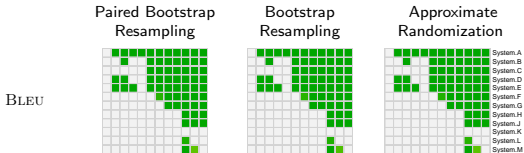
based on the Wilcoxon rank-sum test
- Score each translation sample based on:
  - ① BLEU [Papineni et al., 2002]
  - ② NIST [NIST, 2002]
  - ③ TER [Snover et al., 2005]
  - ④ METEOR [Banerjee and Lavie, 2005]

## Reference Results (ES-EN)

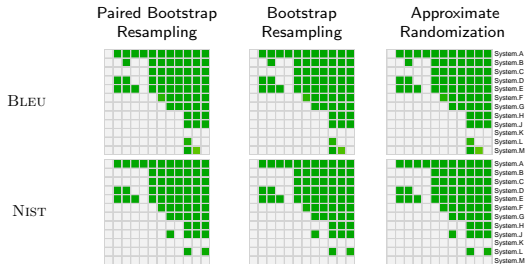
- System comparison based on the segment-level human assessments (ES-EN):



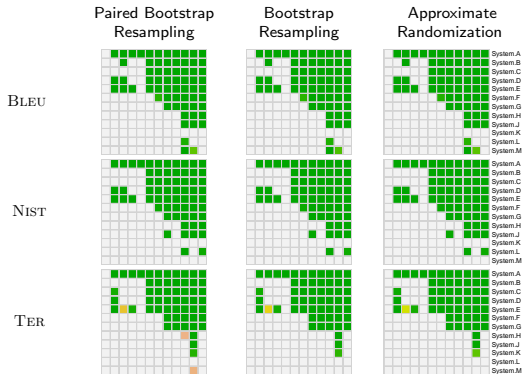
## Pairwise Significance Tests (ES-EN)



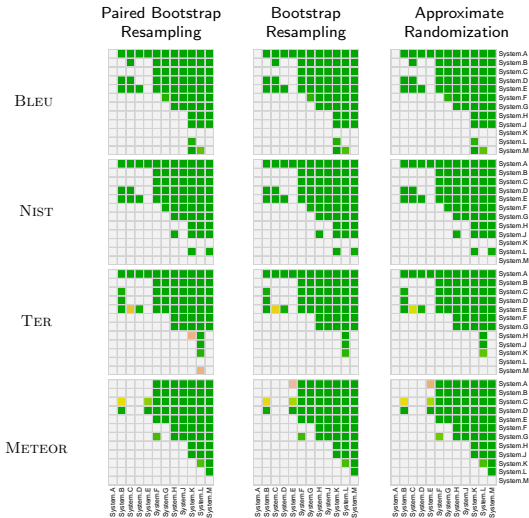
## Pairwise Significance Tests (ES-EN)



## Pairwise Significance Tests (ES-EN)



## Pairwise Significance Tests (ES-EN)





# Accuracy (%) for ES-EN

$p$		Paired Bootstrap	Bootstrap	Approx. Rand.
0.05	BLEU	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]
	NIST	<b>81.8</b> [70.4, 90.2]	<b>81.8</b> [70.4, 90.2]	<b>81.8</b> [70.4, 90.2]
	TER	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]
	METEOR	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]

# Accuracy (%) for ES-EN

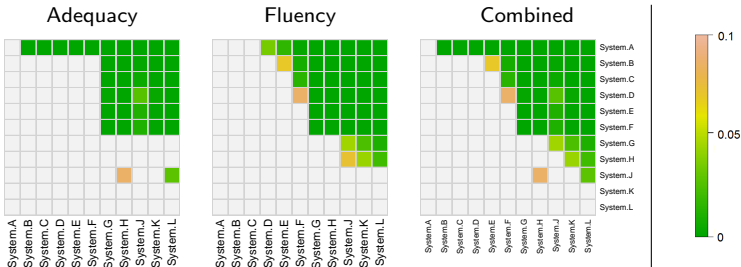
$p$		Paired Bootstrap	Bootstrap	Approx. Rand.
0.05	BLEU	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]
	NIST	<b>81.8</b> [70.4, 90.2]	<b>81.8</b> [70.4, 90.2]	<b>81.8</b> [70.4, 90.2]
	TER	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]
	METEOR	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]
0.01	BLEU	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]
	NIST	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]
	TER	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]
	METEOR	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]

# Accuracy (%) for ES-EN

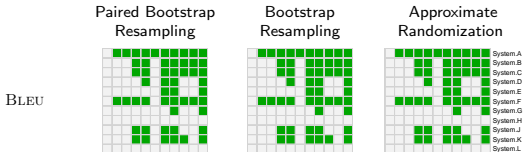
$p$		Paired Bootstrap	Bootstrap	Approx. Rand.
0.05	BLEU	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]
	NIST	<b>81.8</b> [70.4, 90.2]	<b>81.8</b> [70.4, 90.2]	<b>81.8</b> [70.4, 90.2]
	TER	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]
	METEOR	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]
0.01	BLEU	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]
	NIST	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]
	TER	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]
	METEOR	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]
0.001	BLEU	72.7 [60.4, 83.0]	72.7 [60.4, 83.0]	72.7 [60.4, 83.0]
	NIST	72.7 [60.4, 83.0]	72.7 [60.4, 83.0]	72.7 [60.4, 83.0]
	TER	75.8 [63.6, 85.5]	77.3 [65.3, 86.7]	78.8 [67.0, 87.9]
	METEOR	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]	78.8 [67.0, 87.9]

# Human Assessment Ranking (EN-ES)

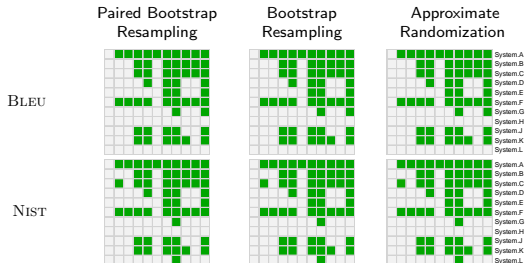
System comparison based on human assessments (EN-ES):



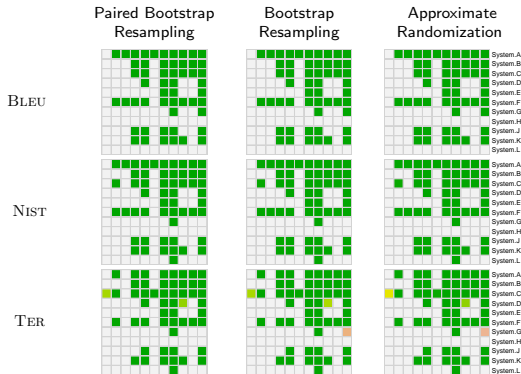
## Pairwise Significance Tests (EN-ES)



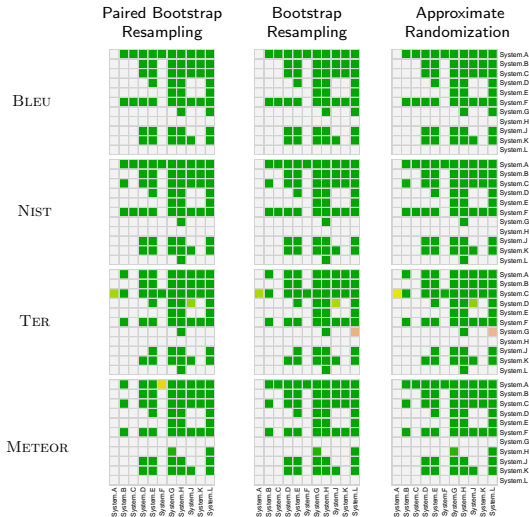
## Pairwise Significance Tests (EN-ES)



## Pairwise Significance Tests (EN-ES)



## Pairwise Significance Tests (EN-ES)





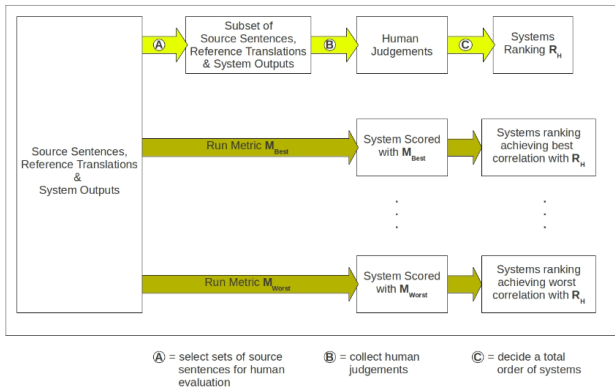
## MT Significance Test Summary

- Very little difference between the three significance tests for either grouping of systems/language pair
- Differences between MT evaluation metrics, but within metric, very little difference across tests
- In terms of agreement with the human evaluations at  $p < 0.05$ :
  - for ES–EN, NIST the most accurate (82% agreement)
  - for EN–ES, BLEU the most accurate (62%(!) agreement)

# Lecture Outline

- ① Automatic Evaluation
- ② Human Evaluation
- ③ MT System Significance Tests
- ④ MT Metric Evaluation
- ⑤ MT Quality Estimation Evaluation

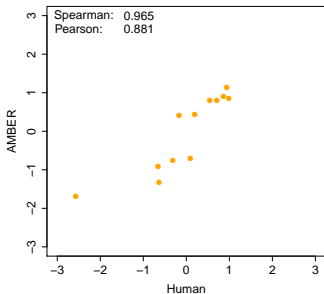
# Document-level Metric Evaluation



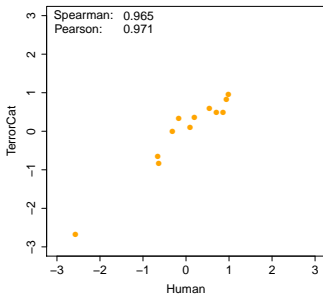
- Metrics are evaluated by how well their scores correlate with human assessment scores

## Correlation with Human Assessment

AMBER



TERRORCAT



- Pearson is unit-free but more sensitive to differences in metric performance than Spearman correlation, useful for evaluation of metrics!

## Significance Tests for MT Metrics

Tests carried out should be for the significance of a *difference* in correlation with human assessment

- Significance of individual correlations with human assessment – not the correct test!
- Randomized methods highly unlikely to be accurate as they don't model the dependent nature of the data.

Appropriate test to use for this purpose is Williams test:

- Test for a *difference* in dependent correlations – ideal for many NLP tasks but not well known

<https://github.com/ygraham/nlp-williams>



# Segment-level Metric Evaluation I

## Segment-level MT Metric Evaluation

- Evaluated by correlation with human assessment (segment-level scores)
- Comparison with WMT rel. pref. judgments been shown to under-reward metrics
- Combination of 15 crowd-sourced continuous rating human judgments into mean scores provides a more accurate gold standard [Graham et al., 2015a]

## Segment-level metrics are more challenging to develop

- Aggregation over the translations in doc. level eval. cancels out much of the random error in individual translation scores – cannot be done in the same way on the segment level
- Subsequently state-of-the-art in seg-level metrics substantially lower than doc. level metrics

## Segment-level Metric Evaluation II

- Best metrics achieve correlation of around 0.5 compared to doc-level of commonly above 0.9 (for MT into English)



# Lecture Outline

- ① Automatic Evaluation
- ② Human Evaluation
- ③ MT System Significance Tests
- ④ MT Metric Evaluation
- ⑤ MT Quality Estimation Evaluation

# Machine Translation Quality Estimation

MT Quality Estimation: the automatic prediction of machine translation quality without the use of reference translations.

**Kein Mensch geht heute noch freiwillig in eine dieser Old-Style-Bibliotheken, oder?**

**No-one goes into one of these old-style libraries voluntarily nowadays, or am I wrong?**

**No one is still in one of these old-style libraries, right?**

**No one goes even today voluntarily to one of these Old style libraries, or?**

Automatically predict the quality of machine-translated text.

# Comparison Measures

Evaluation of quality estimation systems commonly by:

- Pearson Correlation
- Mean Absolute Error
- Root Mean Squared Error

Gold Standards:

- Compare either discrete gold with discrete prediction
- ... or continuous gold with continuous prediction
- Continuous with discrete
- HTER scores
- Post-edit times
- Post-edit rates

## Comparison Measures

Evaluation of quality estimation systems commonly by:

- Pearson Correlation
- Mean Absolute Error [Graham, 2015]
- Root Mean Squared Error [Graham, 2015]

Gold Standards:

- Compare either discrete gold with discrete prediction
- ... or continuous gold with continuous prediction
- ~~continuous with discrete~~ [Graham, 2015]
- HTER scores
- ~~Post-edit times~~ [Graham, 2015]
- Post-edit rates

# Mean Absolute Error

**comparison**

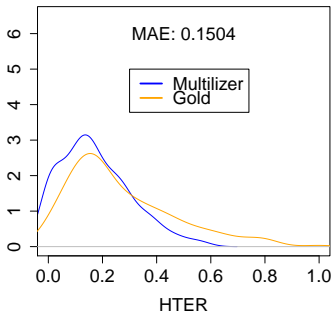
Translation	Gold Label	System Prediction
1	0.91	0.82
2	0.73	0.71
3	0.65	0.51
4	0.54	0.64
5	0.82	0.99

Mean Absolute Error (MAE) in QE: the average absolute difference between gold labels and system predictions computed for a test set of translations.

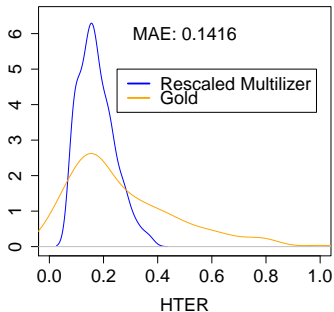
- Lower MAE better

# Evaluation Issues for Widely-used Measures I

(a) Original Predictions



(b) Rescaled Predictions



MAE can be lowered (improved), not only by achieving individual predictions closer to gold labels, but by prediction of aggregate statistics specific to the particular test set gold label distribution.

## Evaluation Issues for Widely-used Measures II

Problems with ability to boost apparent performance by prediction of gold label aggregates:

- Aggregates of gold distribution far easier to predict than individual gold labels
- Aggregates are specific to the choice of test set: very likely to over-estimate the ability of systems to predict the quality of new unseen translations
- Discourages systems from attempting to accurately predict translations in the tails of gold distributions (very good or very bad quality translations) – probably quite important (Moreau & Vogel, 2014)

## Evaluation Issues for Widely-used Measures III

WMT-14 EN-ES Task 1.2: MAE of *all systems* reduced by rescaling according to gold distribution aggregates

	Original MAE	Rescaled MAE
FBK-UPV-UEDIN-wp	0.129	0.125
DCU-rtm-svr	0.134	0.127
USHEFF	0.136	0.133
DCU-rtm-tree	0.140	0.129
DFKI-svr	0.143	0.132
FBK-UPV-UEDIN-nowp	0.144	0.137
SHEFF-lite-sparse	0.150	0.141
Mutilizer	0.150	0.135
baseline	0.152	0.149
DFKI-svr-xdata	0.161	0.146
SHEFF-lite	0.182	0.168



## Evaluation Issues for Widely-used Measures IV

Similar issues arise for other measures that are not unit-free, such as Root Mean Squared Error (RMSE) (W<sub>MT-14</sub> EN-ES Task 1.2)

	Original RMSE	Rescaled RMSE
FBK-UPV-UEDIN-wp	0.167	0.166
DCU-rtm-svr	0.167	0.165
DCU-rtm-tree	0.175	0.169
DFKI-svr	0.177	0.171
USHEFF	0.178	0.178
FBK-UPV-UEDIN-nowp	0.181	0.180
SHEFF-lite-sparse	0.184	0.179
baseline	0.195	0.194
DFKI-svr-xdata	0.195	0.187
Multilizer	0.209	0.181
SHEFF-lite	0.234	0.216

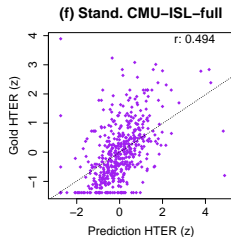
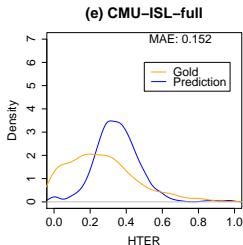
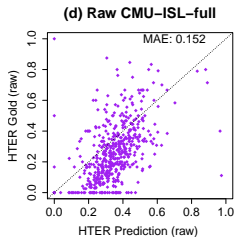
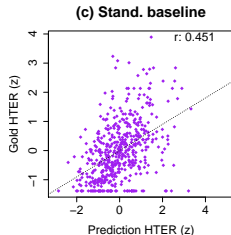
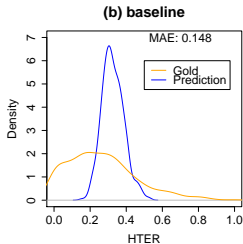
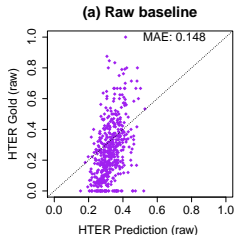
# Pearson Correlation for QE Evaluation I

Pearson's  $r$ : linear correlation between system predictions and gold labels

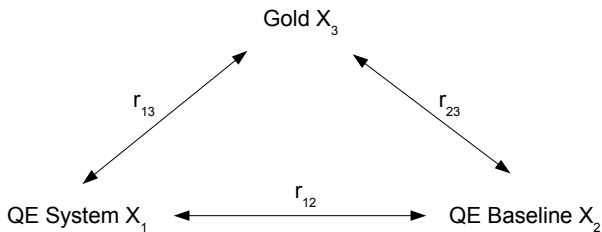
Overcomes problems with MAE and RMSE:

- **Unit-free measure** with a key property being that the correlation coefficient is invariant to separate changes in location and scale in either of the two variables
- Cannot be altered by shifting or rescaling prediction score distributions according to aggregates specific to the test set
- Williams test, theoretical statistical significance test, can be directly applied to test if an increase in correlation is likely to have occurred by chance

# Pearson Correlation for QE Evaluation II



## QE Evaluation Significance Test



Williams Test: Significance of  $r_{13} - r_{23}$

- Test for significance of a *difference* in dependent correlation
- Null Hypothesis:  $r_{13}$  equals  $r_{23}$
- If  $r_{13} > 0$  and  $r_{23} > 0$ , then  $r_{12}$  must also be  $> 0$
- Power of Williams test increased with stronger  $r_{12}$

## Example: WMT-14 PER Prediction I

Training Labels	QE System	$r$	MAE	Original Rank
HTER	A	0.529	—	
PET	B	0.472	0.972	4
HTER	C	0.452	—	
HTER	D	0.444	—	
HTER	E	0.444	—	
HTER	F	0.442	—	
HTER	G	0.441	—	
PET	H	0.430	0.932	2
PET	I	0.423	1.012	8
HTER	J	0.412	—	
PET	K	0.394	1.358	9
PET	L	0.394	1.359	10
PET	M	0.365	0.915	1
HTER	N	0.361	—	
PET	O	0.337	0.951	6
PET	P	0.323	0.940	3
PET	Q	0.288	0.993	7
HTER	R	0.286	—	
HTER	S	0.277	—	
PET	T	0.271	0.972	4
HTER	U	0.011	—	



# References I

- S. Banerjee and A. Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgements. In *Proc. Wkshp. Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–73, Ann Arbor, MI, 2005.
- O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, MA, 2014.
- C. Callison-Burch, M. Osborne, and P. Koehn. Re-evaluating the role of BLEU in machine translation research. In *Proc. 11th Conf. European Chapter of the ACL*, pages 249–256, Trento, Italy, April 2006. ACL.
- M. Denkowski and A. Lavie. Choosing the right evaluation for machine translation: An examination of annotator and automatic metric performance on human judgement tasks. In *Proc. 9th Conf. Assoc. Machine Translation in the Americas*, Denver, CO, 2010.
- George Doddington. Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics. In *Proc. Human Languages Technologies Conf.*, San Diego, California, 2002.
- Y. Graham, T. Baldwin, A. Moffat, and J. Zobel. Continuous measurement scales in human evaluation of machine translation. In *Proc. 7th Linguistic Annotation Wkshp. & Interoperability with Discourse*, pages 33—41, Sofia, Bulgaria, 2013. ACL.
- Y. Graham, T. Baldwin, A. Moffat, and J. Zobel. Is machine translation getting better over time? In *Proceedings of the Fourteenth European Chapter of the Association for Computational Linguistics*, pages 443–51, Gothenburg, Sweden, 2014a.
- Y. Graham, N. Mathur, and T. Baldwin. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–74, Baltimore, MA, 2014b.
- Y. Graham, T. Baldwin, and N. Mathur. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies*, pages 1183–91, Denver, CO, 2015a. ACL.
- Y. Graham, T. Baldwin, A. Moffat, and J. Zobel. Can machine translation systems be evaluated by the crowd alone? In *Journal of Natural Language Engineering*, Cambridge, England, 2015b. Cambridge University.

## References II

- Yvette Graham. Improving evaluation of machine translation quality estimation. In *Proceedings of the Fifty-Third Annual Meeting of the Association for Computational Linguistics*, pages 1804–1813, Beijing, China, 2015. ACL.
- P. Koehn. Statistical significance tests for machine translation evaluation. In *Proc. of Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, 2004. ACL.
- NIST. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Technical report, 2002.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. A Method for Automatic Evaluation of Machine Translation. In *Proc. 40th Ann. Meeting of the Assoc. Computational Linguistics*, pages 311–318, Philadelphia, PA, 2002.
- S. Riezler and J.T. Maxwell. On some pitfalls in automatic evaluation and significance testing for mt. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 57–64, Ann Arbor, MI, 2005. ACL.
- Mathew Snover, Bonnie Dorr, Richard Schwartz, John Makhoul, Linnea Micciula, and Ralph Weischedel. A Study of Translation Error Rate with Targeted Human Annotation. Technical report, College Park, MD, 2005.