Results of the WMT14 Metrics Shared Task

Matouš Macháček   Ondřej Bojar

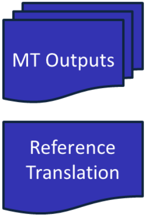
Charles University in Prague

WMT 2014

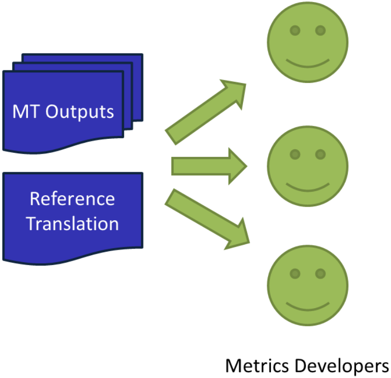
Outline

Metrics Task in a Nutshell

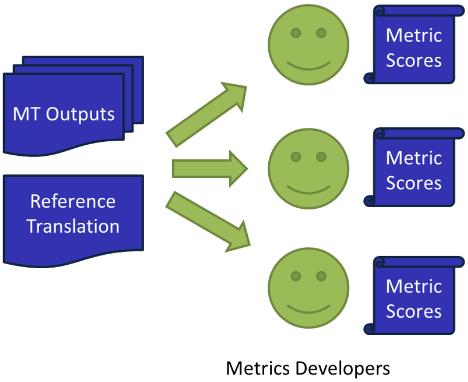
Metrics Task in a Nutshell



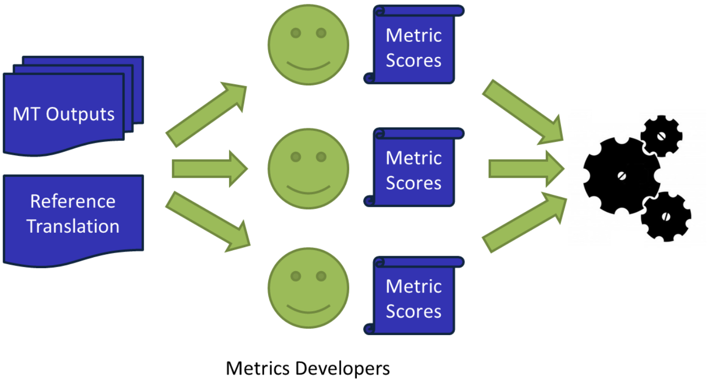
Metrics Task in a Nutshell



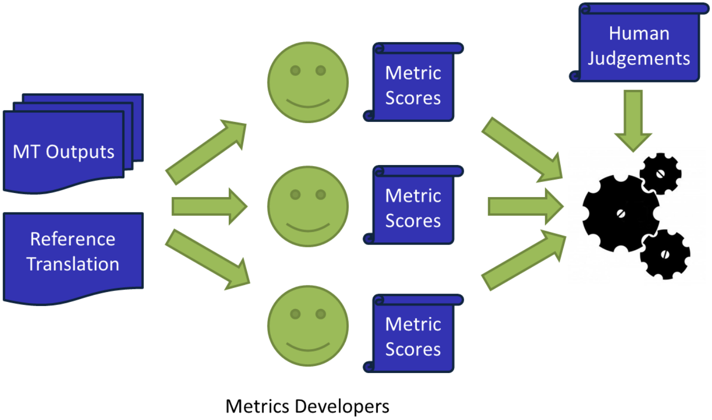
Metrics Task in a Nutshell



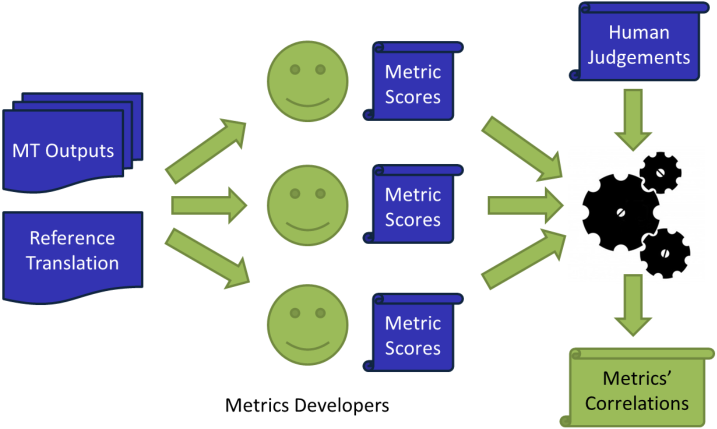
Metrics Task in a Nutshell



Metrics Task in a Nutshell

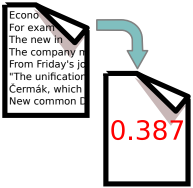


Metrics Task in a Nutshell



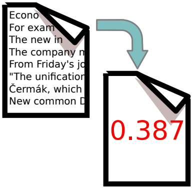
Two subtasks

- System level
  - Participants compute one score for the whole test set, as translated by each of the systems
  - We measure the correlation of these scores with systems' official human scores (TrueSkill)

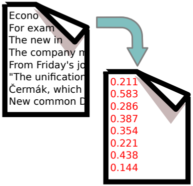


Two subtasks

- System level
  - Participants compute one score for the whole test set, as translated by each of the systems
  - We measure the correlation of these scores with systems' official human scores (TrueSkill)



- Segment level
  - Participants compute one score for each sentence of each system's translation
  - We measure the correlation of these scores with pairwise human judgements



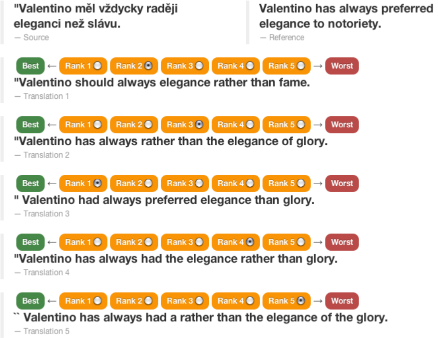
Data

- Data provided to participants:
  - outputs of 110 MT systems involved in the Translation Task
  - 10 reference translations (one for each translation direction)

Data

- Data provided to participants:
  - outputs of 110 MT systems involved in the Translation Task
  - 10 reference translations (one for each translation direction)

- Golden truth
  - Human judgements collected during the evaluation of Translation Task
  - 34,243 “ranking tasks”, this is one of them →
  - Interpreted as 10 pairwise comparisons



Participants and Their Metrics

- 23 metrics from 12 research groups

Metrics	❶	❷	Authors
APAC	✓	✓	Hokkai-Gakuen University (Echizen'ya, 2014)
BEER		✓	University of Amsterdam (Stanojevic and Sima'an, 2014)
RED-*	✓	✓	Dublin City University (Wu and Yu, 2014)
DiscoTK-*	✓	✓	Qatar Computing Research Institute (Guzman et al., 2014)
ELEXR	✓		University of Tehran (Mahmoudi et al., 2013)
LAYERED	✓		Indian Institute of Tech.(Gautam and Bhattacharyya, 2014)
Meteor	✓	✓	Carnegie Mellon University (Denkowski and Lavie, 2014)
AMBER	✓	✓	National Research Council of Canada (Chen and Cherry, 2014)
BLEU-NRC	✓	✓	National Research Council of Canada (Chen and Cherry, 2014)
Parmesan	✓		Charles University in Prague (Barančíková, 2014)
tBLEU	✓		Charles University in Prague (Libovický and Pecina, 2014)
UPC-*	✓	✓	Technical University of Catalunya (Gonzàlez et al., 2014)
VERTa-*	✓	✓	University of Barcelona (Comelles and Atserias, 2014)

❶ – system-level scores

❷ – segment-level scores

Baseline Metrics

- We have computed some metrics as baselines:

Baseline Metrics

- We have computed some metrics as baselines:
- Metrics by `mteval-v13a.pl --international-tokenization`
  - BLEU
  - NIST

Baseline Metrics

- We have computed some metrics as baselines:
- Metrics by `mteval-v13a.pl --international-tokenization`
  - BLEU
  - NIST
- Metrics implemented in Moses Scorer (with moses `tokenizer.perl`)
  - TER
  - WER
  - PER
  - CDER

Computation of System-Level Correlations

- For a given translation direction and metric m...
  - We have human score H_i (TrueSkill) for each MT system s_i
  - We have score of a metric M_i for each MT system s_i

Computation of System-Level Correlations

- For a given translation direction and metric m...
  - We have human score H_i (TrueSkill) for each MT system s_i
  - We have score of a metric M_i for each MT system s_i

- To relate them to each other we use Pearson correlation coefficient:

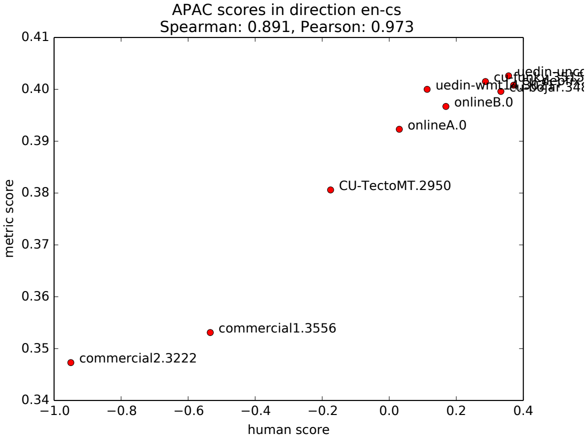
Pearson correlation coefficient (r)

$$r = \frac{\sum_{i=1}^{n}(H_i - \bar{H})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \bar{H})^2}\sqrt{\sum_{i=1}^{n}(M_i - \bar{M})^2}}$$

where $\bar{H}$ and $\bar{M}$ are means of H and M respectively

Why Pearson correlation coefficient

- Spearman's ρ penalizes swapping of similar systems as harsh as swapping very distant systems



- Metrics are likely to behave linearly in the small range of scores

System-Level Correlations into English

From	fr	de	hi	cs	ru	Avg
DiscoTK-party-tuned	.98	**.94**	.96	.97	**.87**	**.94**
LAYERED	.97	.89	**.98**	.94	.85	.93
DiscoTK-party	.97	.92	.86	.98	.86	.92
UPC-STOUT	.97	.91	.90	.95	.84	.91
VERTa-W	.96	.87	.92	.93	.85	.91
VERTa-EQ	.96	.85	.93	.94	.84	.90
tBLEU	.95	.83	.95	.96	.80	.90
BLEU-NRC	.95	.82	.96	.95	.79	.89
BLEU	.95	.83	.96	.91	.79	.89
UPC-IPA	.97	.89	.91	.82	.81	.88
CDER	.95	.82	.83	.97	.80	.87
APAC	.96	.82	.79	.98	.82	.87
REDSys	**.98**	.90	.68	.99	.81	.87
REDSysSent	.98	.91	.64	**.99**	.81	.87
NIST	.96	.81	.78	.98	.80	.87
DiscoTK-light	.96	.93	.56	.95	.79	.84
Meteor	.98	.93	.46	.98	.81	.83
WER	.95	.76	.61	.97	.81	.82
AMBER	.95	.91	.51	.74	.80	.78
ELEXR	.97	.86	.54	.94	-.40	.58

System-Level Correlations out of English

Into	fr	hi	cs	ru	Avg	de[1]
NIST	.94	.98	.98	.93	**.96**	.20
CDER	.95	.95	.98	.94	.95	.28
AMBER	.93	**.99**	.97	.93	.95	.24
METEOR	.94	.98	.98	.92	.95	.26
BLEU	.94	.97	.98	.91	.95	.22
PER	.94	.93	**.99**	**.94**	.95	.19
APAC	.95	.94	.97	.93	.95	.35
TBLEU	.93	.97	.97	.91	.95	.24
BLEU-NRC	.93	.97	.97	.90	.95	.20
ELEXR	.89	.96	.98	.94	.94	.26
TER	.95	.83	.98	.93	.92	.32
WER	**.96**	.52	.98	.93	.85	**.36**
PARMESAN	–	–	.96	–	.96	–
UPC-IPA	.94	–	.97	.92	.94	.28
REDSYSSENT	.94	–	–	–	.94	.21
REDSYS	.94	–	–	–	.94	.21
UPC-STOUT	.94	–	.94	.92	.93	.30

[1] German results are separate because they differ too much

System-Level Correlations Summary

- Overall high correlations
- Best metrics reach 0.99 (different metrics for different language pairs) or .96 (average of the best metric across language pairs)
- Baseline metrics (NIST, CDER, BLEU, PER) surprisingly good out of English
  - ... also WER except for English→Hindi.
- The results into German are very low
  - Probably caused by high number (18) of participating systems
  - It is very difficult for the metrics to discriminate systems of similar quality
- METEOR suffers when evaluating translations from non-Latin script

Computation of Segment-Level Correlations

- A metric is expected to predict the result of the manual pairwise comparison
- The Kendall's τ is used to measure this quality

Computation of Segment-Level Correlations

- A metric is expected to predict the result of the manual pairwise comparison
- The Kendall's τ is used to measure this quality

The basic formula of Kendall's τ

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|}$$

- *Concordant* – comparisons for which a given metric agrees with human
- *Discordant* – comparisons for which a given metric does not agree
- $\tau \in [-1, 1]$

Computation of Segment-Level Correlations

- A metric is expected to predict the result of the manual pairwise comparison
- The Kendall's τ is used to measure this quality

The basic formula of Kendall's τ

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|}$$

- *Concordant* – comparisons for which a given metric agrees with human
- *Discordant* – comparisons for which a given metric does not agree
- $\tau \in [-1, 1]$

Example

Human	Metric
A < B	A < B
C > A	C > A
C > B	C < B

$$\tau = \frac{2-1}{2+1} = \frac{1}{3}$$

Computation of Segment-Level Correlations

- A metric is expected to predict the result of the manual pairwise comparison
- The Kendall's τ is used to measure this quality

The basic formula of Kendall's τ

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|}$$

- *Concordant* – comparisons for which a given metric agrees with human
- *Discordant* – comparisons for which a given metric does not agree
- $\tau \in [-1, 1]$

Example

Human	Metric
A < B	A < B
C > A	C > A
C > B	C < B

$$\tau = \frac{2 - 1}{2 + 1} = \frac{1}{3}$$

Computation of Segment-Level Correlations

- A metric is expected to predict the result of the manual pairwise comparison
- The Kendall's τ is used to measure this quality

The basic formula of Kendall's τ

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|}$$

- *Concordant* – comparisons for which a given metric agrees with human
- *Discordant* – comparisons for which a given metric does not agree
- $\tau \in [-1, 1]$

Example

Human	Metric
$A < B$	$A < B$
$C > A$	$C > A$
$C > B$	$C < B$

$$\tau = \frac{2 - 1}{2 + 1} = \frac{1}{3}$$

Computation of Segment-Level Correlations

- A metric is expected to predict the result of the manual pairwise comparison
- The Kendall's τ is used to measure this quality

The basic formula of Kendall's τ

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|}$$

- *Concordant* – comparisons for which a given metric agrees with human
- *Discordant* – comparisons for which a given metric does not agree
- $\tau \in [-1, 1]$

Example

Human	Metric
A $<$ B	A $<$ B
C $>$ A	C $>$ A
C $>$ B	C $<$ B

$$\tau = \frac{2-1}{2+1} = \frac{1}{3}$$

- What to do with A $=$ B?

Generalization of the Kendall's τ formula

WMT12 variant

		Metric		
		$<$	$=$	$>$
Human	$<$	1	-1	-1
	$=$	X	X	X
	$>$	-1	-1	1

- The table specifies coefficients in the numerator of Kendall's τ
  - 1 corresponds to *Concordant*
  - -1 corresponds to *Discordant*
- The coefficient in denominator is always 1, except for X

Generalization of the Kendall's τ formula

WMT12 variant

		Metric $<$	Metric $=$	Metric $>$
Human	$<$	1	-1	-1
Human	$=$	X	X	X
Human	$>$	-1	-1	1

Example

Human	Metric
$A < B$	$A < B$
$A < B$	$A < B$
$A > B$	$A = B$
$A = B$	$A > B$

$$\tau = \frac{2 \cdot 1 + 1 \cdot (-1)}{2 + 1}$$

- The table specifies coefficients in the numerator of Kendall's τ
  - 1 corresponds to *Concordant*
  - -1 corresponds to *Discordant*
- The coefficient in denominator is always 1, except for X

Generalization of the Kendall's τ formula

WMT12 variant

		Metric		
		$<$	$=$	$>$
Human	$<$	1	-1	-1
	$=$	X	X	X
	$>$	-1	-1	1

Example

Human	Metric
$A < B$	$A < B$
$A < B$	$A < B$
$A > B$	$A = B$
$A = B$	$A > B$

$$\tau = \frac{2 \cdot 1 + 1 \cdot (-1)}{2 + 1}$$

- The table specifies coefficients in the numerator of Kendall's τ
  - 1 corresponds to *Concordant*
  - -1 corresponds to *Discordant*
- The coefficient in denominator is always 1, except for X

Generalization of the Kendall's τ formula

WMT12 variant

		Metric <	Metric =	Metric >
Human	<	1	-1	-1
Human	=	X	X	X
Human	>	-1	-1	1

Example

Human	Metric
$A < B$	$A < B$
$A < B$	$A < B$
$A > B$	$A = B$
$A = B$	$A > B$

$$\tau = \frac{2 \cdot 1 + 1 \cdot (-1)}{2 + 1}$$

- The table specifies coefficients in the numerator of Kendall's τ
  - 1 corresponds to *Concordant*
  - -1 corresponds to *Discordant*
- The coefficient in denominator is always 1, except for X

Generalization of the Kendall's τ formula

WMT12 variant

		Metric $<$	Metric $=$	Metric $>$
Human	$<$	1	-1	-1
Human	$=$	X	X	X
Human	$>$	-1	-1	1

Example

Human	Metric
$A < B$	$A < B$
$A < B$	$A < B$
$A > B$	$A = B$
$A = B$	$A > B$

$$\tau = \frac{2 \cdot 1 + 1 \cdot (-1)}{2 + 1}$$

- The table specifies coefficients in the numerator of Kendall's τ
  - 1 corresponds to *Concordant*
  - -1 corresponds to *Discordant*
- The coefficient in denominator is always 1, except for X

Generalization of the Kendall's τ formula

WMT12 variant

		Metric		
		$<$	$=$	$>$
Human	$<$	1	-1	-1
	$=$	X	X	X
	$>$	-1	-1	1

Example

Human	Metric
$A < B$	$A < B$
$A < B$	$A < B$
$A > B$	$A = B$
$A = B$	$A > B$

$$\tau = \frac{2 \cdot 1 + 1 \cdot (-1)}{2 + 1}$$

- The table specifies coefficients in the numerator of Kendall's τ
  - 1 corresponds to *Concordant*
  - -1 corresponds to *Discordant*
- The coefficient in denominator is always 1, except for X
- Why should metric's ties be penalized as discordant?

More variants of the Kendall's τ formula

WMT13 variant

		Metric		
		$<$	$=$	$>$
Human	$<$	1	X	-1
	$=$	X	X	X
	$>$	-1	X	1

- In WMT13, we excluded metrics ties like the human ties. (X items are not considered at all)
- It turned out that it allows gaming by assigning a lot of ties, which lowers the denominator.

More variants of the Kendall's τ formula

WMT13 variant

		Metric $<$	Metric $=$	Metric $>$
Human	$<$	1	X	-1
Human	$=$	X	X	X
Human	$>$	-1	X	1

- In WMT13, we excluded metrics ties like the human ties. (X items are not considered at all)
- It turned out that it allows gaming by assigning a lot of ties, which lowers the denominator.

WMT14 variant

		Metric $<$	Metric $=$	Metric $>$
Human	$<$	1	0	-1
Human	$=$	X	X	X
Human	$>$	-1	0	1

- In WMT14 we return the count of metric ties into the denominator, so a metric which often yields ties gets lower score

Segment-Level Correlations into English: Kendall's τ

	WMT14 variant						WMT var.	
From	**fr**	**de**	**hi**	**cs**	**ru**	**Avg**	**12**	**13**
DiscoTK-party-tuned	**.43**	**.38**	.43	**.33**	**.35**	**.39**	**.39**	.39
BEER	.42	.34	**.44**	.28	.33	.36	.36	.36
REDcombSent	.41	.34	.42	.28	.34	.36	.35	.36
REDcombSysSent	.41	.34	.42	.28	.34	.36	.35	.36
Meteor	.41	.33	.42	.28	.33	.35	.34	.36
REDSysSent	.40	.34	.39	.28	.32	.35	.33	.35
REDSent	.40	.34	.38	.28	.32	.35	.33	.35
UPC-IPA	.41	.34	.37	.27	.32	.34	.34	.34
UPC-STOUT	.40	.34	.35	.28	.32	.34	.34	.34
VERTa-W	.40	.32	.39	.26	.31	.34	.32	.34
VERTa-EQ	.41	.31	.38	.26	.31	.34	.32	.34
DiscoTK-party	.39	.33	.36	.26	.31	.33	.33	.33
AMBER	.37	.31	.36	.25	.29	.32	.30	.32
BLEU-NRC	.38	.27	.32	.23	.27	.29	.27	.30
sentBLEU	.38	.27	.30	.21	.26	.29	.26	.29
APAC	.36	.27	.29	.20	.28	.28	.24	.29
DiscoTK-light	.31	.22	.24	.19	.21	.23	.23	.23
DiscoTK-light-kool	.00	.00	.00	.00	.00	.00	-1.00	**.68**

Segment-Level Correlations out of English: Kendall's τ

	WMT14 variant						WMT var.	
Into	fr	de	hi	cs	ru	Avg	12	13
BEER	.29	**.27**	.25	**.34**	**.44**	**.32**	**.31**	**.32**
METEOR	.28	.24	.26	.32	.43	.31	.28	.31
AMBER	.26	.23	**.29**	.30	.40	.30	.27	.30
BLEU-NRC	.26	.20	.23	.30	.39	.28	.24	.29
APAC	.25	.21	.20	.29	.39	.27	.22	.28
SENTBLEU	.26	.19	.23	.29	.38	.27	.23	.28
UPC-STOUT	.28	.23	–	.28	.42	.30	.30	.31
UPC-IPA	.26	.23	–	.30	.43	.30	.29	.31
REDSENT	**.29**	.24	–	–	–	.27	.25	.27
REDCOMBSYSSENT	.29	.24	–	–	–	.27	.25	.27
REDCOMBSENT	.29	.24	–	–	–	.27	.25	.27
REDSYSSENT	.29	.24	–	–	–	.26	.23	.27

Overall Summary

- Metrics task still interesting! (12 teams took part.)
- (But the results are hard to interpret.)
- System-level correlations in the 0.8 – 1.0 range
- Segment-level still poor: around 0.4

Overall Summary

- Metrics task still interesting! (12 teams took part.)
- (But the results are hard to interpret.)
- System-level correlations in the 0.8 – 1.0 range
- Segment-level still poor: around 0.4

Chef's tips for evaluation:

- System-level
  - into English: DiscoTK-party-tuned, LAYERED, UPC-STOUT
  - out of English: NIST, CDER, AMBER
- Segment-level
  - into English: DiscoTK-party-tuned, BEER, REDcombSent
  - out of English: BEER, Meteor, AMBER

References I

Petra Barančíková. 2014. Parmesan: Improving Meteor by More Fine-grained Paraphrasing. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

Boxing Chen and Colin Cherry. 2014. A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

Elisabet Comelles and Jordi Atserias. 2014. VERTa participation in the WMT14 Metrics Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

Hiroshi Echizen'ya. 2014. Application of Prize based on Sentence Length in Chunk-based Automatic Evaluation of Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

References II

Shubham Gautam and Pushpak Bhattacharyya. 2014. LAYERED: Description of Metric for Machine Translation Evaluation in WMT14 Metrics Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

Meritxell Gonzàlez, Alberto Barrón-Cedeño, and Lluís Màrquez. 2014. IPA and STOUT: Leveraging Linguistic and Source-based Features for Machine Translation Evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

Francisco Guzman, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. DiscoTK: Using Discourse Structure for Machine Translation Evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

Jindřich Libovický and Pavel Pecina. 2014. Tolerant BLEU: a Submission to the WMT14 Metrics Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

References III

Alireza Mahmoudi, Heshaam Faili, MohammadHossein Dehghan, and Jalal Maleki. 2013. ELEXR: Automatic Evaluation of Machine Translation Using Lexical Relationships. In Félix Castro, Alexander Gelbukh, and Miguel González, editors, *Advances in Artificial Intelligence and Its Applications*, volume 8265 of *Lecture Notes in Computer Science*, pages 394–405. Springer Berlin Heidelberg.

Milos Stanojevic and Khalil Sima'an. 2014. BEER: A Smooth Sentence Level Evaluation Metric with Rich Ingredients. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

Xiaofeng Wu and Hui Yu. 2014. RED, The DCU Submission of Metrics Tasks. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.