A Cluster-Based Representation for Multi-System MT Evaluation

Nicolas Stroppa and Karolina Owczarzak
National Centre for Language Technology
School of Computing, Dublin City University

Overview

National Centre for Language Technology

- Evaluating Machine Translation (MT)
  - automatic metrics
  - human judgement
  - "My MT is better than yours": unreliability of system rankings

- The need for statistical significance
  - bootstrap
  - approximate randomization

- Cluster representation
  - "My MT might not be better than yours, but it's definitely better than his": groupings and confidence levels

- Automatic metrics vs. human judgement on the cluster level: cluster comparison

Automatic metrics in MT evaluation

- Fast and cheap way to evaluate Machine Translation quality
- Used for system development or cross-system comparison
- Most popular: BLEU, NIST, GTM, METEOR
- Criticism of string-level comparison and inadequate correlations with human judgement
- Small differences in automatic scores between systems due to chance: data type, missing punctuation, unknown word, weather, butterfly flapping its wings in Ecuador
- Hard rankings of systems based on raw evaluation results not advisable
- Statistical significance testing necessary

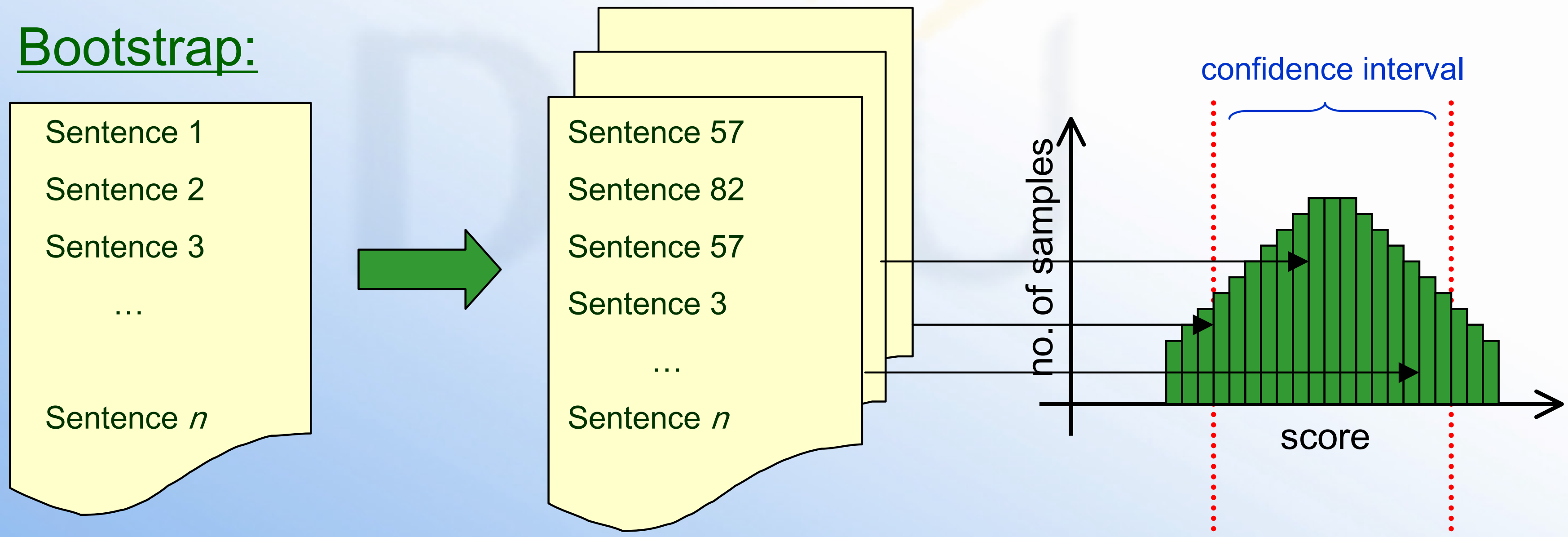
Humans in MT Evaluation

- Slow and expensive way to evaluate Machine Translation quality
- Used in shared tasks (ACL SMT workshop 2007)
- Standard scale: Adequacy 1-5, Fluency 1-5
- Standard frame of reference for developing automatic metrics
- Human evaluation not so consistent either:
  - inter-annotator K ~0.23
  - intra-annotator K ~0.5 (Callison-Burch et al. 2007)
- Small differences in human scores between systems due to chance: personal writing style preferences, imperfect knowledge or understanding, tiredness, distraction, the fact that it's Tuesday – humans are unreliable and inconsistent! (I, for one, welcome our new AI overlords)
- Hard rankings of systems based on human evaluation results not advisable
- Statistical significance testing necessary

Statistical Significance Testing

- Null hypothesis: two MT systems are of the same quality
- Difference between their scores only significant if statistical evidence against null hypothesis
- Significance testing for MT evaluation: non-parametric methods
  - bootstrap (Efron and Tibshirani 1993, Koehn 2004)
  - approximate randomization (Noreen 1989, Riezler and Maxwell 2005)

Bootstrap:



Approximate randomization

- More appropriate to MT eval (Riezler and Maxwell 2005; Collins et al. 2005)

MT_1	MT_2		MT'_1	MT'_2
MT_1 sentence 1	MT_2 sentence 1	→	MT_1 sentence 1	MT_2 sentence 1
MT_1 sentence 2	MT_2 sentence 2		MT_2 sentence 2	MT_1 sentence 2
MT_1 sentence 3	MT_2 sentence 3		MT_1 sentence 3	MT_2 sentence 3
…	…		…	…
MT_1 sent n	MT_2 sent n		MT_2 sent n	MT_1 sent n

IF $MT_1 > MT_2$ **THEN** $MT_1 > MT'_1$

$MT_2 < MT'_2$

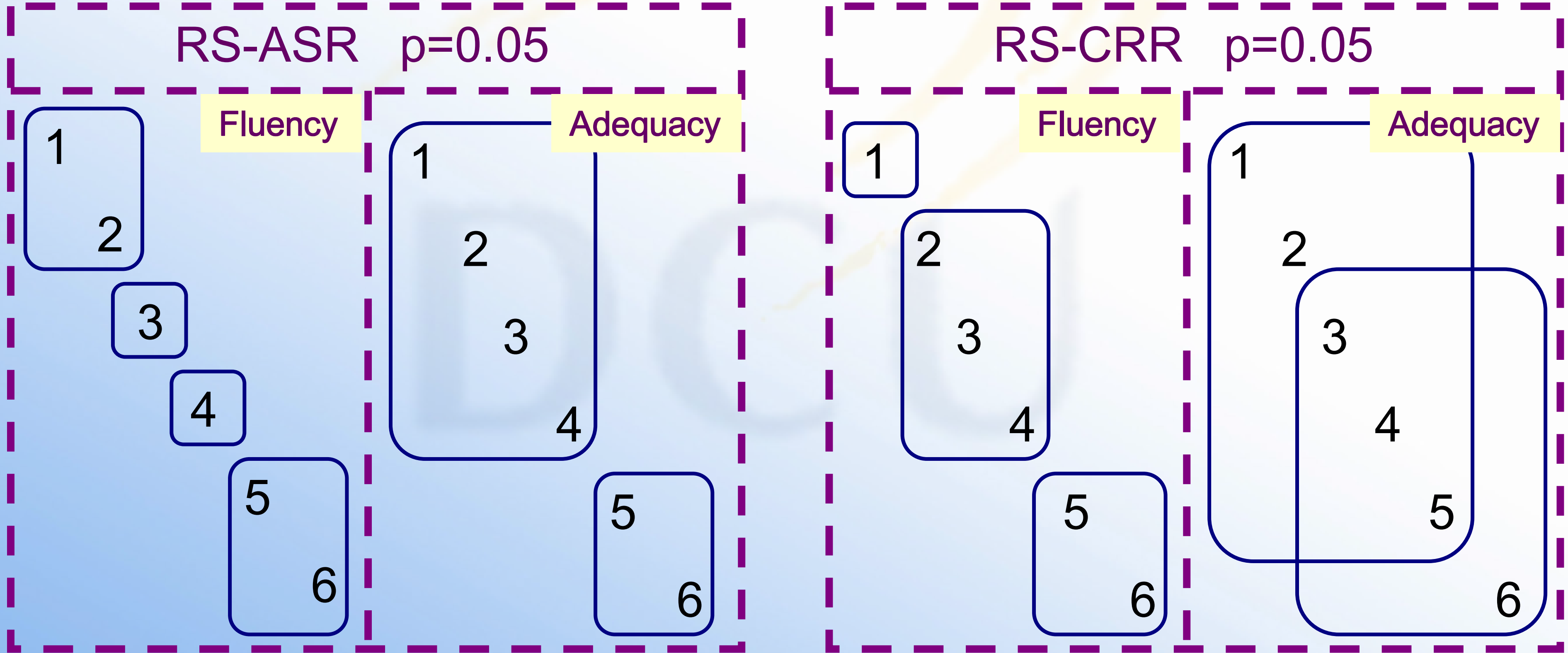
$MT_1 - MT_2 > MT'_1 - MT'_2$

$$\mathbf{p} = \frac{\left(\sum_{i=1}^{k} v_i\right) + 1}{k + 1}$$

Cluster-based representation

National Centre for Language Technology

- Approximate randomization likely to show some MT systems cannot be distinguished (at a certain confidence level)
- Clusters contain MT systems that are pairwise indistinguishable
- Clusters can overlap: A !> B, B !> C, A > C



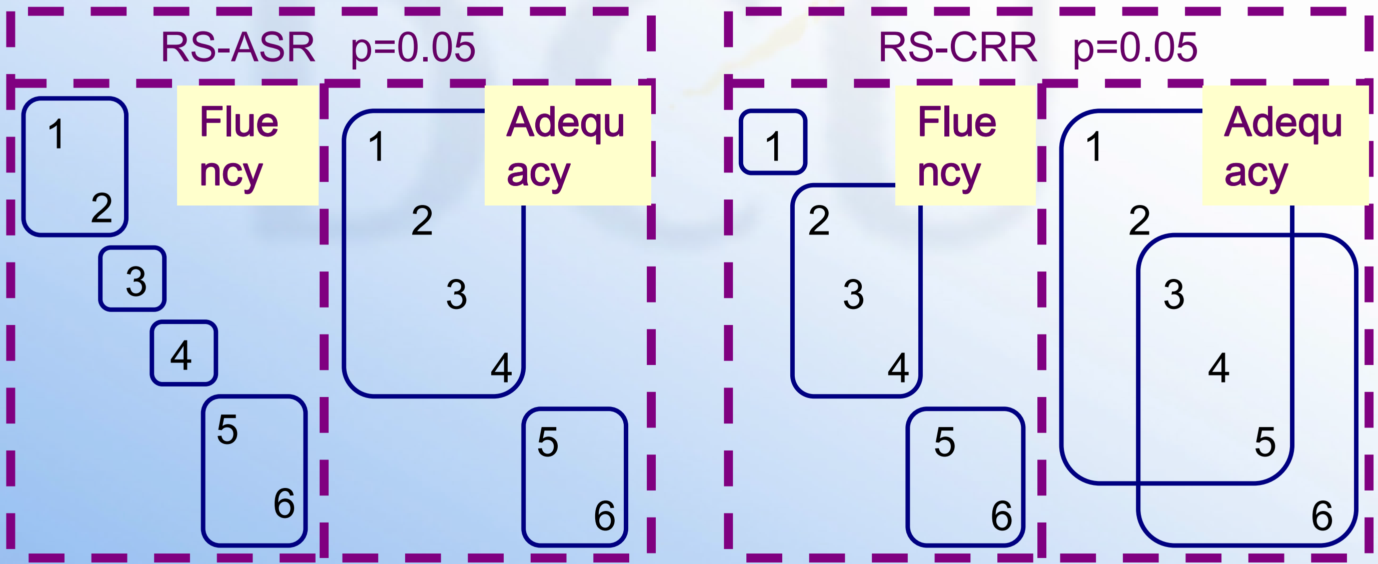
Comparing clusters

- Adaptation of the Rand statistics (Haldiki et al. 2001)

- Compare relationships of *pairs of MT systems* across cluster rankings

$$\text{score(rel1,rel2)} = \begin{cases} 1 & \text{if (rel1 = rel2)} \\ -1 & \text{if (rel1 = `<<' and rel2 = `>>')} \\ -1 & \text{if (rel1 = `>>' and rel2 = `<<')} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{score(ranking1,ranking2)} = \frac{2 * \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{score}(C(i,j), D(i,j))}{n * (n-1)}$$

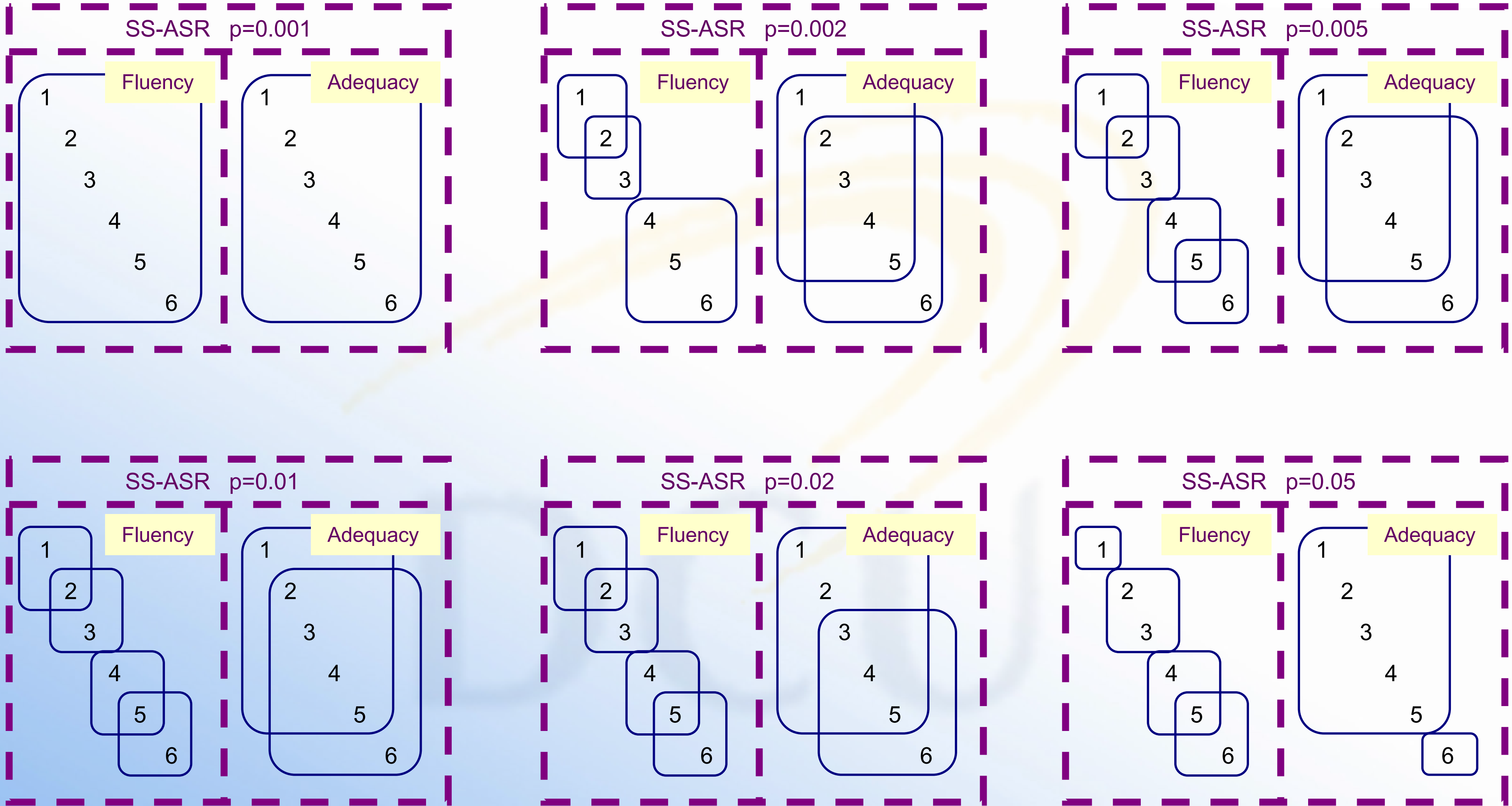


Experiment – clusters and comparisons

National Centre for Language Technology

- Data: IWSLT 2006 Chinese-English translations
  - 500 segments
  - six MT systems
  - three conditions: spontaneous speech (SS-ASR), read speech with automatic speech recognition (RS-ASR), read speech with correct recognition (RS-CRR)
  - human evaluation (adequacy and fluency) for all translations
  - evaluated with BLEU, NIST, GTM, METEOR
- Approximate randomization on all scorings
  - varying confidence levels (p=0.001, p=0.002, p=0.005, p=0.01, p=0.02, p=0.05)
  - analysis of resulting clusters
- Comparison of clusters based on human and automatic scores
- Comparison of clusters based on different automatic scores
- Relationship between confidence level and human – automatic correlation

Clusters and confidence levels



Comparison of human and automatic clusters

p = 0.05

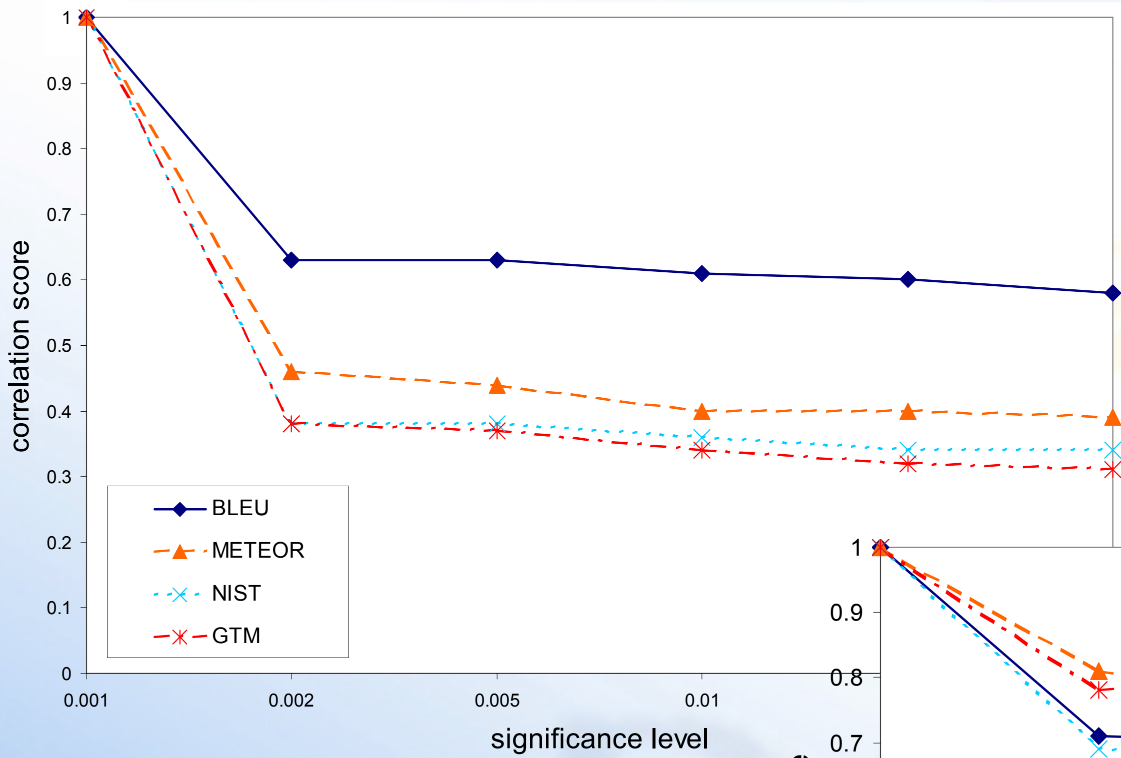
		Fluency	Adequacy
SS-ASR	BLEU	0.47	0.4
	NIST	0	0.6
	METEOR	0	0.53
	GTM	-0.13	0.6
RS-ASR	BLEU	0.47	0.33
	NIST	0.4	0.27
	METEOR	0.33	0.13
	GTM	0.2	0.2
RS-CRR	BLEU	0.73	0.47
	NIST	0.4	0.27
	METEOR	0.53	0.26
	GTM	0.33	0.33
Mixed Track	BLEU	0.58	0.7
	NIST	0.34	0.64
	METEOR	0.39	0.71
	GTM	0.31	0.7

Comparing automatic metrics (Mixed Track)

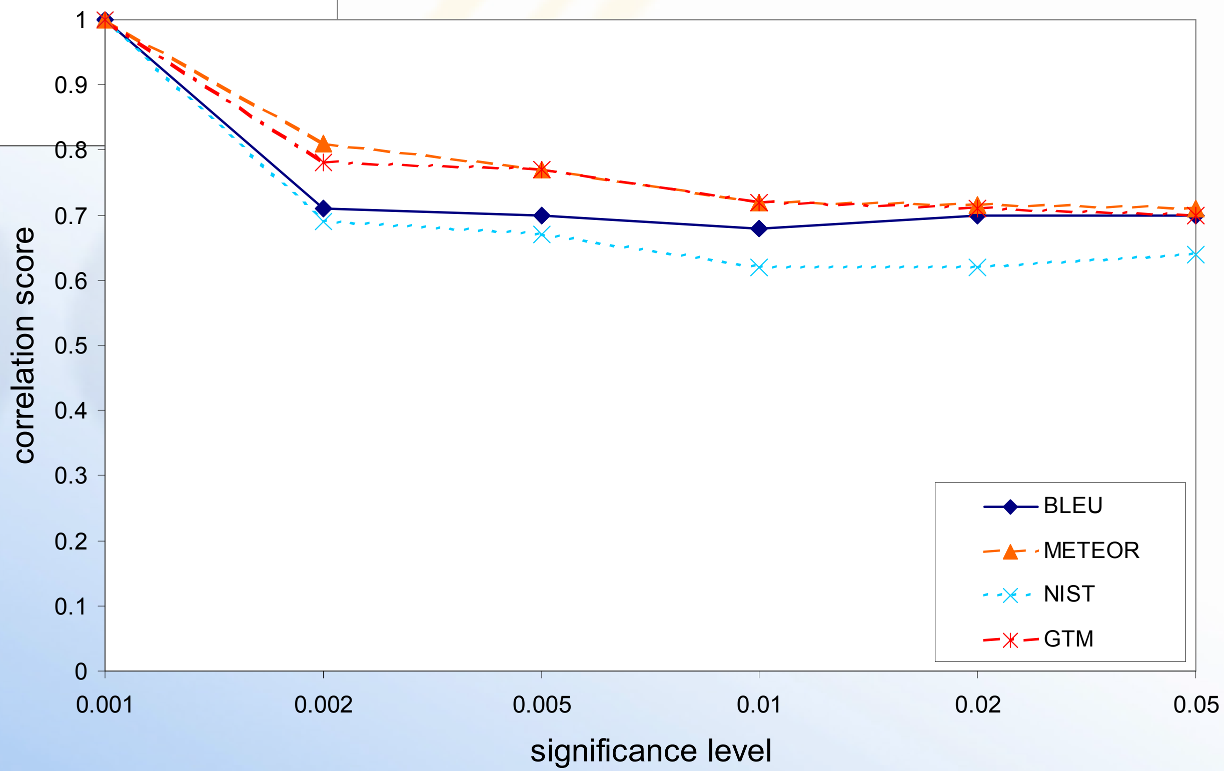
p = 0.05

	BLEU	NIST	METEOR
NIST	0.64	-	-
METEOR	0.77	0.79	-
GTM	0.7	0.79	0.86

Correlations and confidence levels



Correlation (human scores of **fluency**, automatic metrics) vs. significance level



Correlation (human scores of **adequacy**, automatic metrics) vs. significance level

Discussion and conclusions

- Small differences in (human or automatic) scores may be accidental
- Statistical significance testing necessary for Truth and Justice (and A Hard-Boiled Egg)
- Produce clusters of MT systems at given significance level
- Trade-off: as level of required confidence increases, it‘s more difficult to distinguish between MT systems
- Cluster comparison – another method for comparison of system-level human and automatic scores
- Evaluating automatic metrics necessary at both system and segment level
  - metrics with high system-level correlations good for multiple MT system comparisons (shared tasks etc.)
  - metrics with high segment-level correlations good for MT development
- Automatic metrics cannot reflect well fluency and adequacy at the same time

References

National Centre for Language Technology

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrisic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the ACL 2007 Workshop on Statistical Machine Translation*, pages 136-158, Prague, Czech Republic.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL 2005*, pages 531–540, Ann Arbor, MI.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of HLT 2002*, pages 128–132, San Diego, CA.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2004. Precision and recall of machine translation. In *Proceedings of HLT-NAACL 2003*, volume 2, pages 61–63, Edmonton, Canada.

Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley-Interscience, New York, NY.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, PA.

Stefan Riezler and John Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the the ACL Workshop on Intrisic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 57–64, Ann Arbor, MI.