MetricsMaTr10
Evaluation Overview & Summary of Results

Kay Peterson & Mark Przybocki
Brian Antonishek, Mehmet Yilmaz, Martial Michel

MetricsMaTr10

- NIST Metrics for Machine Translation Challenge

A research challenge to improve MT metrology

- Development of ***Intuitive*** metrics
- Development of metrics that provide ***Insights*** into ***quality***

- Partnered with WMT
  - A single evaluation
  - Larger data sets – releasable data
  - Greater exposure

MetricsMaTr10 (continued)

- Second MetricsMaTr evaluation
  - In 2008, **13** participants submitted **32** metrics
  - In 2010, **14** participants submitted **26** metrics
- Schedule:

Begin date	*End date*	*task*
January 11		Announcement of evaluation plans
March 26	May 14	Metric submission
May 15	June/July	Metric installation and data set scoring
July 2		Preliminary release of results
July 15	July 16	Workshop
September		Official results posted on NIST web space

SUBMITTED METRICS

14 MetricsMaTr10 Participants

Affiliation	*URL*	*Metric name(s)*		
Aalto University of S&T *		MT-NCD	MT-mNCD	
BabbleQuest	**http://www.babblequest.com/badger2**	**badger-2.0-lite**	**badger-2.0-full**	
City University of Hong Kong *	**http://mega.ctl.cityu.edu.hk/ctbwong/ATEC**	**ATEC-2.1**		
Carnegie Mellon *	**http://www.cs.cmu.edu/~alavie/METEOR**	**meteor-next-rank**	meteor-next-hter	meteor-next-adq
Columbia University	**http://www1.ccls.columbia.edu/~SEPIA**	**SEPIA**		
Charles University Prague *		SemPOS	SemPOS-BLEU	
Dublin City University *		DCU-LFG		
University of Edinburgh *		LRKB4	LRHB4	
Harbin Institute of Technology		i-letter-BLEU	i-letter-recall	**SVM-rank**
National University of Singapore *	**http://nlp.comp.nus.edu.sg/software**	TESLA	TESLA-M	
Stanford University NLP		Stanford		
University of Maryland	**http://www.umiacs.umd.edu/~snover/terp**	**TERp**		
University Politecnica de Catalunya & University of Barcelona *	http://www.lsi.upc.edu/~nlp/Asiya	**IQmt-Drdoc**	**IQmt-DR**	**Iqmt-ULCh**
University of Southern California, ISI	**http://www.isi.edu/publications/licensed-sw/BE/index.html**	**BEwT-E**	Bkars	

BLUE entries participated in MetricsMaTr08

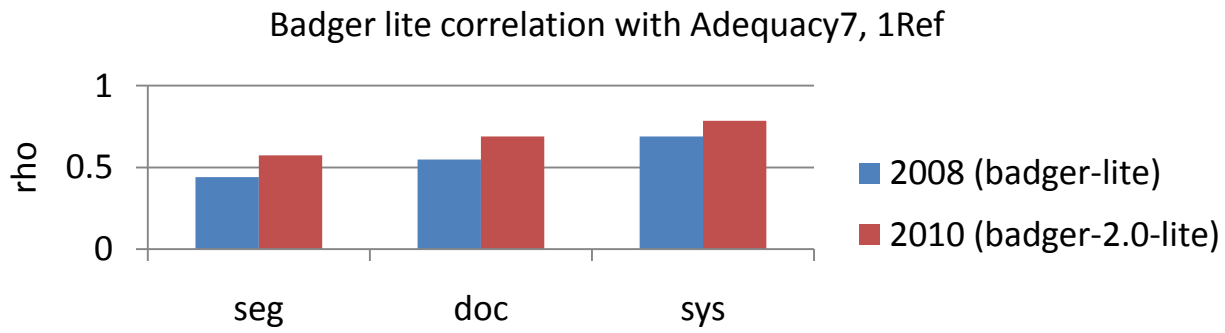
* Represented with a paper in ACL 2010 main or WMT/MetricsMaTr workshop proceedings

Aalto University of S&T

Metric:	**MT-NCD**
Features:	-base on “Normalized Compression Distance (NCD) -works on the character level -otherwise works similarly to most other MT evaluation metrics
Metric:	**MT-mNCD**
Features:	-enhancements include flexible word matching through stemming and WordNet synsets (English) -analogously to MaTr-08 entries: M-BLEU and M-TER -borrows from METEOR: aligner module -aligned words in the reference are replaced by their counterparts -score is then calculated between the two -multiple references treated individually, (unclear: best score?)

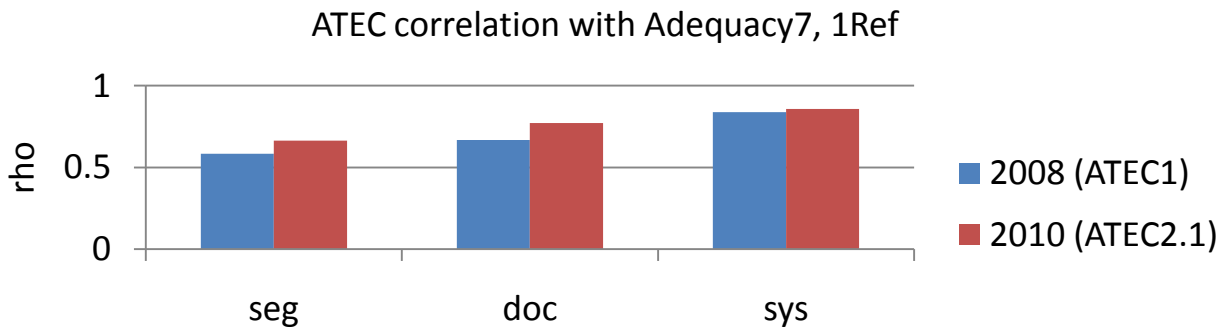
BabbleQuest

Metric:	**badger-2.0-full**
Features:	-employs “SimMetrics” by Sam Chapman at Sheffield University -contains a normalization knowledgebase for all 2010 challenge languages -Uses Smith Waterman Gotoh similarity measure (similar to Levenshtein)
Metric:	**badger-2.0-lite**
Features:	-does not perform word normalization



City University of Hong Kong

Metric:	**ATEC-2.1**
Features:	-parameters optimized for word choice and word order -use Porter stemmer and WordNet for stems and synonym matches -uses WordNet-based measure of word similarity for word matches -matches are weighted by “informativeness” -uses position distance, order distanced and phrase size (word order)



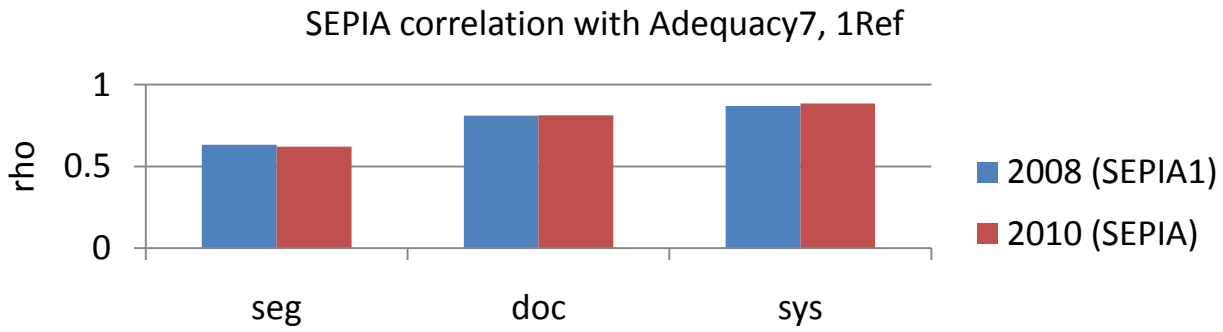
Carnegie Mellon

Metric:	**meteor-next-rank**
Features:	-meteor-next calculates a similarity score based on exact, stem, synonym, and paraphrase matches -"rank" is tuned to maximize rank consistency on human ranking of WMT09
Metric:	**meteor-next-hter**
Features:	-"hter" is tuned to segment-level length-weighted Pearson's correlation with GALE P2 HTER data
Metric:	**meteor-next-adq**
Features:	-"adq" is tuned to segment-level length-weighted Pearson's correlation with NIST OpenMT 2009 human adequacy judgments

Consistent high correlation

Columbia University

Metric:	**SEPIA**
Features:	-Precision-based, syntactically aware evaluation metric -Assigns bigger weights to grammatical structured bigrams with long surface spans -Uses a dependency representation for both hypotheses and reference(s) -Configurable for different combinations of: structural n-grams, surface n-grams, POS tags, or dependency relations and lemmatization



Charles University Prague

Metric:	**SemPOS**
Features:	-Computes the overlap of content bearing word lemmas between the hyp and ref translation given a fine-grained semantic part-of-speech (sempos) -Outputs average overlapping score across all sempos types
Metric:	**SemPOS-BLEU**
Features:	-linear combination of SemPos and BLEU BLEU is calculated on surface forms only autosemantic words

Dublin City University

Metric:	**DCU-LFG**
Features:	-dependency-based metric -produces 1-best LFG dependencies and allow triple matches where labels differ -sorts matches according to match level and dependency type; weighted to maximize correlation with human judgment -final match is the sum of weighted matches

University of Edinburgh

Metric:	**LRscore (LRKB4, LRHB4)**
Features:	-Measures reordering success using permutation distance metrics -The reordering component is combined with the lexical metric -Language independent

Harbin Institute of Technology

Metric:	**i-letter-BLEU**
Features:	-Normal BLEU based on letters -Maximum length N-gram is average length for each sentence
Metric:	**i-letter-recall**
Features	-Geometric mean of N-gram recall based on letters -Maximum length N-gram is average length for each sentence
Metric:	**SVM-rank**
Features:	-Uses support vector machine rank models to predict ordering of system translations -Features include: Meteor-exact, BLEU-cum-(1,2,5), BLEU-ind-(1,2), ROUGE-L recall, letter-based TER, letter-based BLEU-cum-5, letter-based ROUGE-L recall, and letter-based ROUGE-S recall.

National University of Singapore

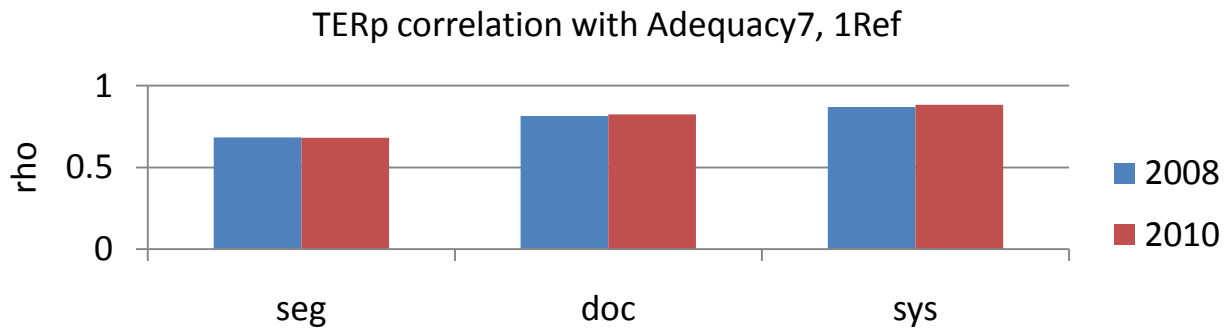
Metric:	**TESLA-M**
Features:	-Based on matching n-grams (1-3) with the use of WordNet synonyms -Discounts function words
Metric:	**TESLA**
Features:	-TESLA-M plus the use of bilingual phrase tables for phrase-level synonyms -Feature weights tuned with SVM-rank over development data

Stanford University NLP

Metric:	**Stanford**
Features:	-String edit distance metric with multiple similarity matching techniques -The model represents a conditional random field

University of Maryland

Metric:	**TERp**
Features:	-Extends TER by using stemming, synonymy, and paraphrasing -Accepts tunable costs -Adds a brevity and length penalty



University Politecnica de Catalunya & University of Barcelona

Metric:	**ULCH**
Features:	-Arithmetic mean over a heuristically-defined set of metrics
Metric:	**DR**
Features:	-Arithmetic mean over a set of three metrics based on discourse representations operating at the segment level ** respectively computing lexical overlap ** morphosyntactic overlap ** semantic tree matching
Metric:	**DRdoc**
Features:	"DR" at the whole document level

Note: Better correlation with WMT than MetricsMaTr tests

University of Southern California, ISI

Metric:	**BEwT-E**
Features:	-A recall-oriented metric -Compares “basic elements (Bes)” between two translations -”Bes’” are content words and various combinations of syntactically-related words -Is English specific
Metric:	**Bkars**
Features:	-Produces a score both with and without stemming ** Uses the Snowball package of stemmers -Is NOT English specific

Bkars consistently in Top 10 (seg, doc, sys Adequacy7)

Baseline Metrics

- All MetricsMaTr08 entries
- Focus on BLEU (-c = case-sensitive)
  - MT-EVAL version 11b (MetricsMaTr08)
  - MT-EVAL version 12 (MetricsMatr08 non-English)
  - ✓**MT-EVAL version 13a (OpenMT09)**
- NIST (-c = case-sensitive)

Baseline Metrics

Metric:	BLEU-v11b
Version:	MTEVAL version 11b
Description:	Modified BLEU-4 with an improved brevity penalty Case-sensitive N-gram co-occurrence statistics Official metric of recent NIST Open MT evaluations

Metric:	BLEU-v12
Authoring Affiliation:	NIST (IBM) (2008)
Description:	Updated BLEU-v11b (above) with UTF-8 tokenization rules for non-English target languages

Metric:	BLEU-v13a
Authoring Affiliation:	NIST (IBM) (2009)
Description:	XML version Command line options for some Non-English translations

MetricsMaTr08: Workshop Suggestions

- Data sets – 100% XML (*yes*)
- Include a stress test of the data (*somewhat*)
  - Installation included a "check set" (empty segments)
  - Long segments (NA)
- Archival of results, process, metrics (*yes*)
  - Online scores
  - Special Issue of MT Journal
- Allow more time for running metrics (*no*)
  - Metrics are becoming more complex (installation and operation)

EVALUATION DATA

Important Note about the Eval Data

- MetricsMaTr data is not publicly available
  1. We do not have permission to release the system translations
  2. Some data is to be used (reused) in future MT technology evaluations
  3. Some data required NIST to sign a license agreement for its inclusion
  4. This eval data will be reused in future MetricsMaTr evaluations
  5. The GALE subset of the data will likely be released via LDC in the future

Evaluation Data Set Specifics

Primary

Origin	Source	Target	Genre(s)	Doc. count	Segment count	Words (est.)	Systems (mt+ht)	Refs. available
MT08	Arabic	English	NW, WB	42	405	15,100	10 + 2	4
	Chinese	English	NW, WB	51	607	15,000	10 + 2	4
GALE P2	Arabic	English	NW, WB	45	469	11,450	3	1
	Chinese	English	NW, WB	47	392	10,150	3	1
GALE P2.5	Arabic	English	BN	20	210	5,300	2	1
	Chinese	English	BC, BN	42	289	10,000	3	1
TRANSTAC Jan07	Arabic	English	Dialog	15	433	5,150	5 + 2	4
TRANSTAC Jul07	Arabic	English	Dialog	47	419	6,450	5 + 2	4
	Farsi	English	Dialog	25	414	4,550	5 + 2	4

Evaluation Data Set Specifics

Secondary

Origin	Source	Target	Genre(s)	Doc. count	Segment count	Words (est.)	Systems (mt+ht)	Refs. available
CESTA run1	Arabic	French	General	16	298	27,950	(2 + 1)	4
	English	French	General	15	790	21,350	(5 + 1)	4
CESTA run2	Arabic	French	Health	30	824	20,100	(1 + 1)	4
	English	French	Health	16	917	22,550	(5 + 1)	4
TRANSTAC Jan07	English	Arabic	Dialogs				5	4

- European Language Resources Association provided CESTA data (ELRA catalog reference E0020, http://catalog.elra.info/product_info.php?products_id=994)
  - General:
    - official journal of the European Community (JOC)
    - the UNESCO conference
  - Health:
    - websites Health Canada, UNICEF, WHO, and FHI

Evaluation Data Set Specifics

WMT

Source	Target	Genre	Documents	Segments	Words (est.)	Systems (single + combo)	References
Czech	English	NW	94	2034	42,000	7+5	1
French	English				54,000	16+8	
German	English				49,000	18+7	
Spanish	English				52,000	10+4	
English	Czech				50,000 each	12+5	
English	French					15+4	
English	German					14+4	
English	Spanish					12+4	

- Parallel corpus
  - Same data set (docs, segs) for each language pair
  - System combination test subset of WMT10 test set

MetricsMaTr-Provided Development Data

Data Attributes1	NIST Open MT-06	TRANSTAC
Genre	Newswire	Training dialogs
Number of documents	25	1 (included as sample)
Total number of segments	249	17
Source Language	Arabic	Iraqi Arabic
Number of system translations	8	5

- A sampling of what was to be included in the evaluation data set
- Limited assessment types (adequacy and preference)
- Metric development was not limited to this data

English vs. Foreign Target Language

- All metrics were run on the (3) data sets
  - Primary, secondary, and WMT data
  - If no processing errors, scores are reported
- All metrics were run in appropriate tracks (1Ref, 4Ref)

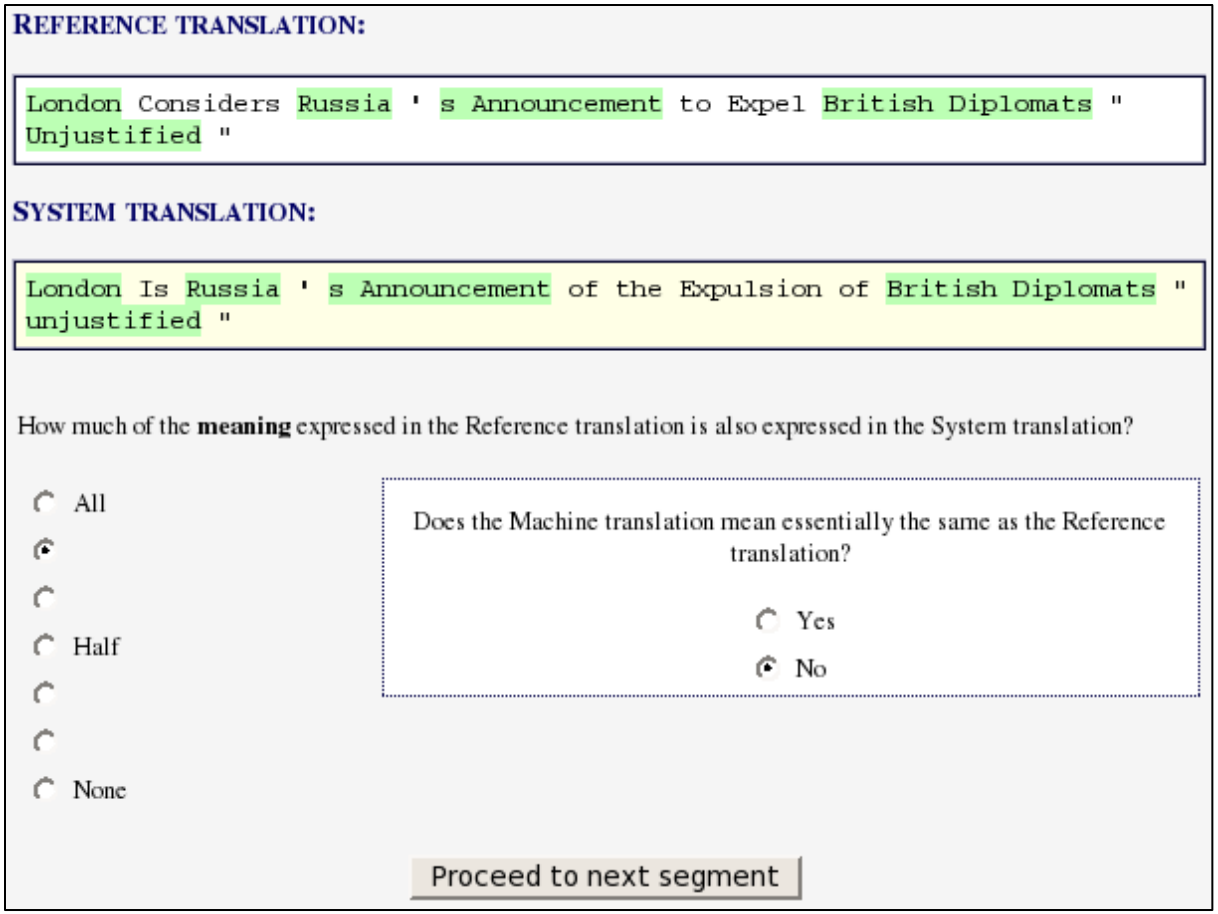
Human Assessment Types

Data subset	Adequacy 7pt	Yes/No decision	Adequacy 5pt	Preference	Fluency 5pt	HTER	Low Level concept	Adequacy 4pt	DLPT*	Relative Rank
MT08	√	√		√					√	
GALE	√	√		√		√				
TRANSTAC	√	√		√			√	√		
CESTA			√		√					
WMT										√

- These types of human assessments will be briefly described
- Most SOURCE documents were reviewed for ILR difficulty (not WMT)
- Adequacy7 + Adequacy Yes/No and Preference were done specifically for the original MetricsMaTr set
  - All other types of assessment were pre-existing and are thus limited to the eval sets they stem from
- Current analysis focuses on Adequacy7

Semantic Adequacy7 and Yes/No

(MT08, GALE, TRANSTAC)



Allowing for 2-off category judgments, we achieve over 90% inter annotator agreement

- Comparison of:
  - 1 reference translation
  - 1 system translation
- Word matches highlighted as a visual aid
- Decision:
  - "Quantitative" (7-point scale)
  - "Qualitative" (Yes/No)
- At least 2 independent judgments for each segment in MetricsMaTr08 test set

MetricsMaTr Data Adequacy7 Score Distribution

- ~54K Independent judgments

Adequacy Score		Coverage	
7 (All)	Yes	20.8%	21.4%
	No	0.6%	
6	Yes	14.9%	21.6%
	No	6.7%	
5	Yes	8.7%	19.0%
	No	10.3%	
4 (Half)	*No*		18.8%
3	*No*		9.2%
2	*No*		5.6%
1 (None)	*No*		4.4%

- ~ 25K Avgs of multiple judgments

Avg. Adequacy Score		Coverage	
6+ to 7	Yes	21.5%	23.9%
	mixed	2.2%	
	No	0.2%	
5+ to 6	Yes	10.2%	22.6%
	mixed	9.0%	
	No	3.4%	
4+ to 5	Yes	1.2%	21.3%
	mixed	9.2%	
	No	10.9%	
3+ to 4	Mixed	2.0%	17.0%
	No	15.0%	
2+ to 3	Mixed	0.1%	9.4%
	No	9.3%	
1+ to 2	*No*	5.8%	5.8%

DLPT* (MT08)

- MT comprehension test
- Test questions developed from source data
- Subjects review MT output and try to answer the questions

Through the MFLTS (Sequoyah) program, this test is being extended to cover multiple language pairs and to increase the size of the test.

Other Assessments

- Preference Judgments (MaTr data)
- 5-pt and 4-pt Adequacy (CESTA, TRANSTAC)
- Traditional 5-pt Fluency (CESTA)
  - Performed prior to Adequacy test
- Concept Transfer (TRANSTAC)
  - Bilingual judges determine in the concepts present in the source data are also present in the resulting translation
- Relative Rank (WMT)

Summary (Data/Human Assessments)

- Many human assessment types in MetricsMaTr
  - Added WMT's Ranking assessment
- Focus for current analysis will remain on Adequacy7 (and some on Adequacy Yes/No, HTER)
- Future:
  - Investigate (better) human assessment types
  - Release some (half?) current MetricsMaTr test set
  - Add MFLTS ILR-based scoring data
  - Add MFLTS expanded DLPT* data
  - Translation Memory Assessment project data

Availability of Results

- Detailed public release on MetricsMaTr10 data: http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2010/results
- Today's talk: Overview of completed analysis
  - Limited to one correlation statistic (Spearman's rho)
  - Limited to target language English data
  - Focus on 1 reference track
  - Focus on MetricsMaTr test set
- Some submitted metrics not included in results due to installation issues
- WMT10 results: http://www.statmt.org/wmt10/results.html

Correlation-Based Rankings

1Ref, Adequacy7, Target Eng, Seg/Doc/Sys

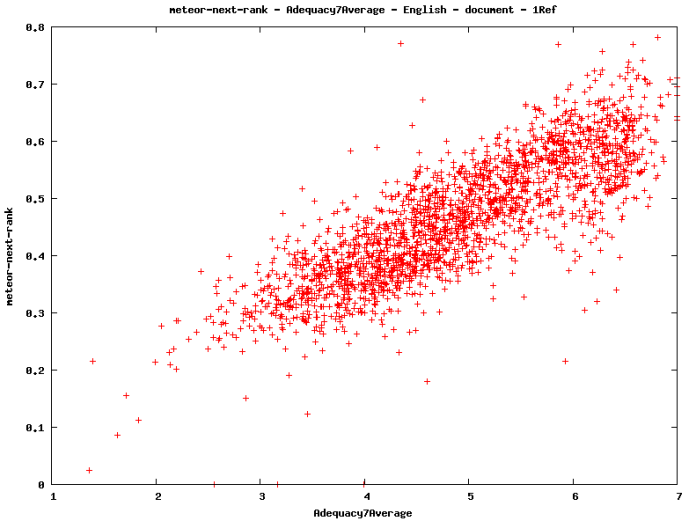
Rank	Seg rho 25473 data points	Doc rho 2179 data points	Sys rho 89 data points
1	meteor-next-rank	meteor-next-rank	meteor-next-rank
2	TERp	meteor-next-adq	meteor-next-adq
3	meteor-next-adq	meteor-next-hter	meteor-next-hter
4	meteor-next-hter	i-letter-recall	i-letter-recall
5	ATEC-2.1	i-letter-BLEU	i-letter-BLEU
6	i-letter-recall	TERp	SEPIA
7	i-letter-BLEU	***NIST-c***	TERp
8	Bkars	SEPIA	***NIST-c***
9	SEPIA	Bkars	Bkars
10	***NIST-c***	***BLEU-4-v13a-c***	DCU-LFG
11	***BLEU-4-v13a-c***	ATEC-2.1	ATEC-2.1
12	badger-2.0-full	DCU-LFG	***BLEU-4-v13a-c***
13	BEwT-E	BEwT-E	BEwT-E
14	badger-2.0-lite	badger-2.0-full	badger-2.0-full
15	DCU-LFG	badger-2.0-lite	badger-2.0-lite
16	TESLA	TESLA	TESLA
17	MT-mNCD	TESLA-M	IQMT-DR
18	MT-NCD	SemPOS-BLEU	TESLA-M
19	SemPOS-BLEU	MT-mNCD	SemPOS-BLEU
20	TESLA-M	IQMT-DR	SemPOS
21	IQMT-DR	IQMT-DRdoc	IQMT-DRdoc
22	SemPOS	SemPOS	MT-mNCD
23	IQMT-DRdoc	MT-NCD	MT-NCD

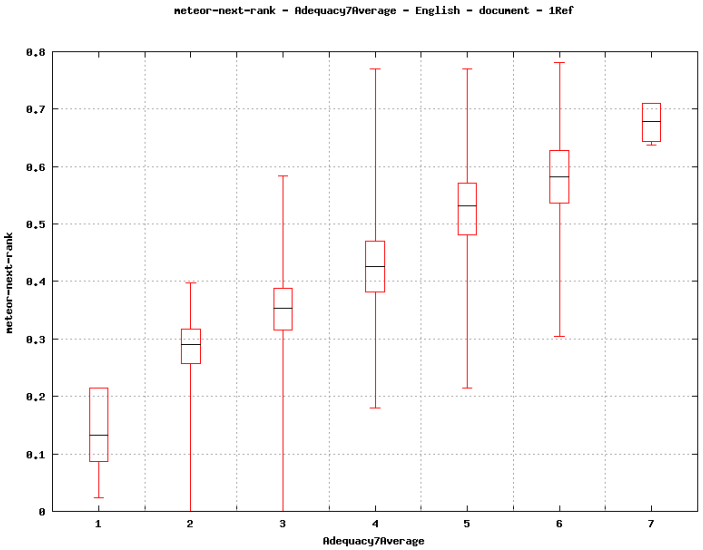
- Ranks based on Spearman's rho correlation

Bold italics
= baseline metrics

Plot Examples

1Ref, Adequacy7, Target Eng, Doc





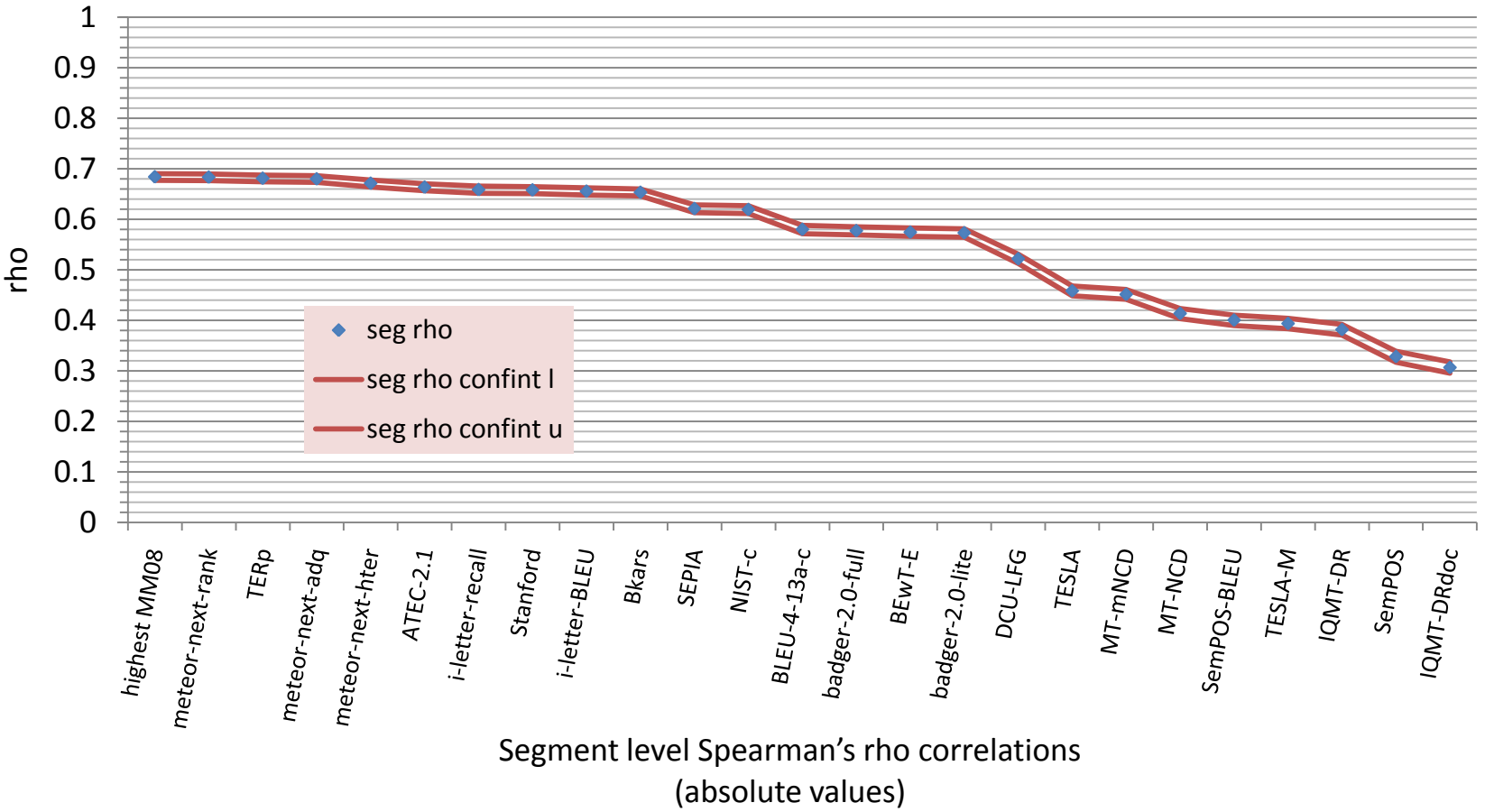
- Scatter and box-and-whiskers plot for one of the strongly correlating metrics
  - Box plot shows metric scores are completely separated for the central 50% of data points at 2-off human assessment bins

Levels of Analysis

- Goal of analysis:
  - Segment level:
    - Investigate low-level metric usefulness
    - Segment level correlations support fine-grained error analysis
  - Document level:
    - Investigate metric usefulness at the “natural” (cohesive one-topic) document level
  - System level:
    - Investigate metric usefulness at system level
    - System level has been the main level under investigation at technology evaluations such as NIST OpenMT

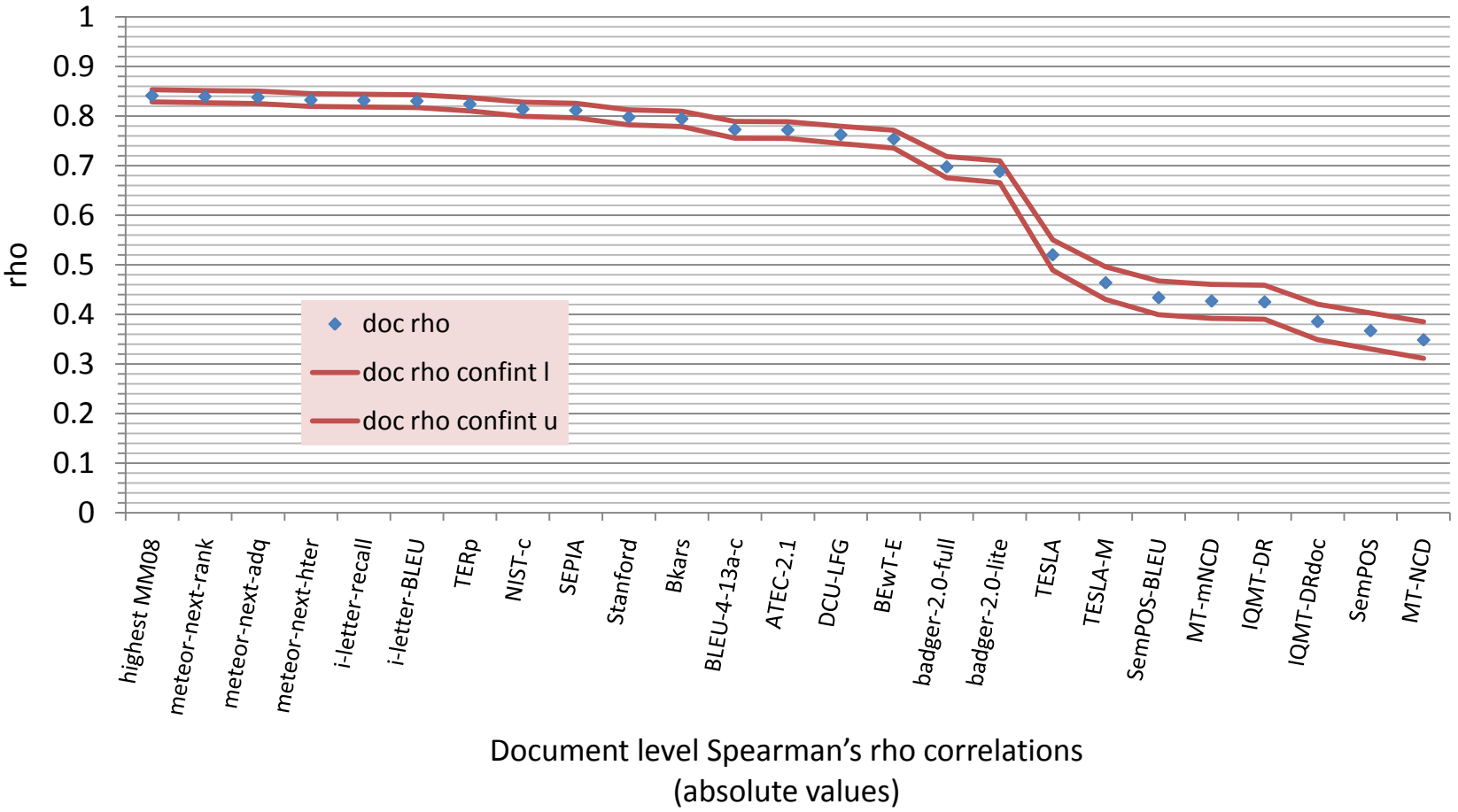
Overall Correlations

1Ref, Adequacy7, Target Eng, Seg



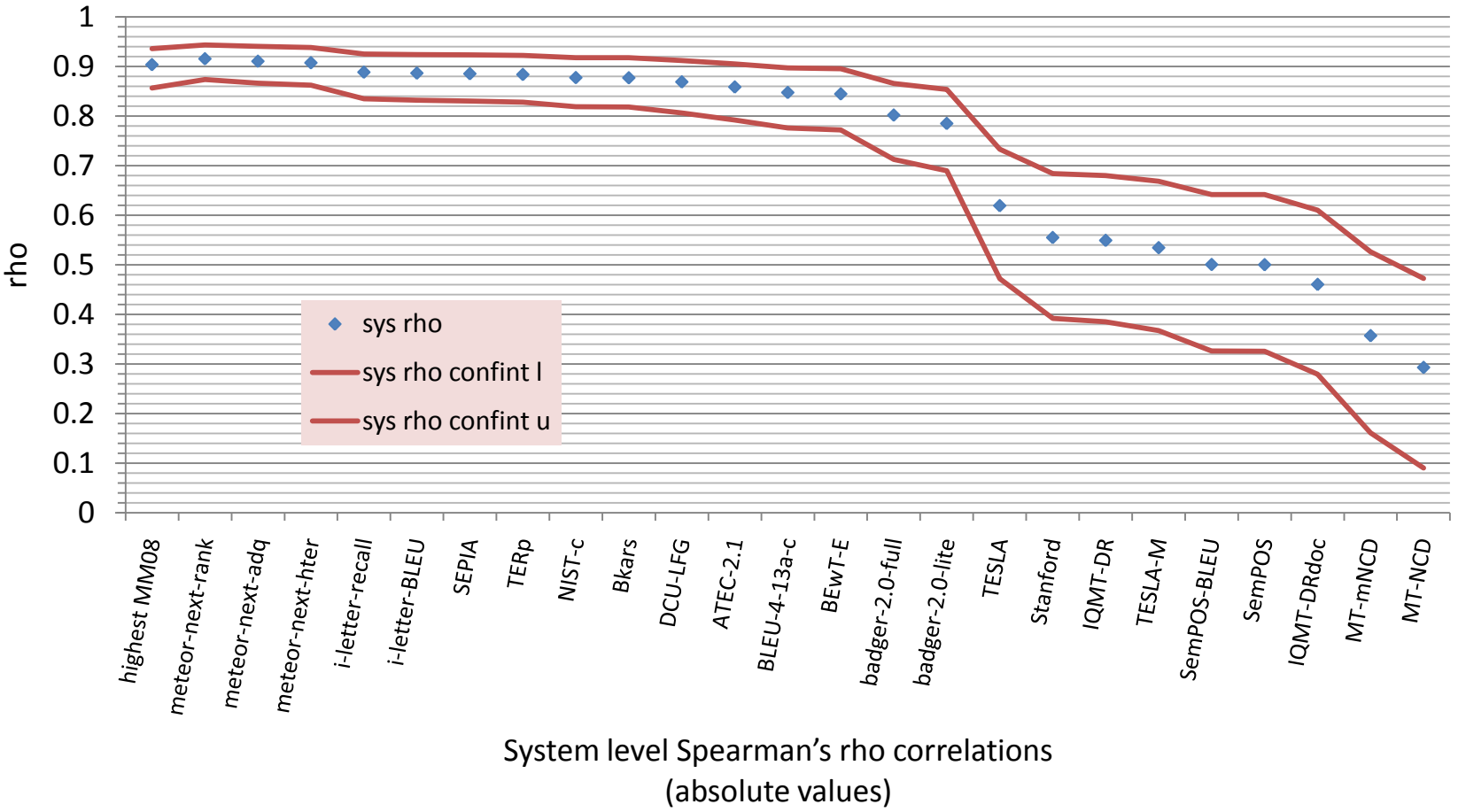
Overall Correlations

1Ref, Adequacy7, Target Eng, Doc



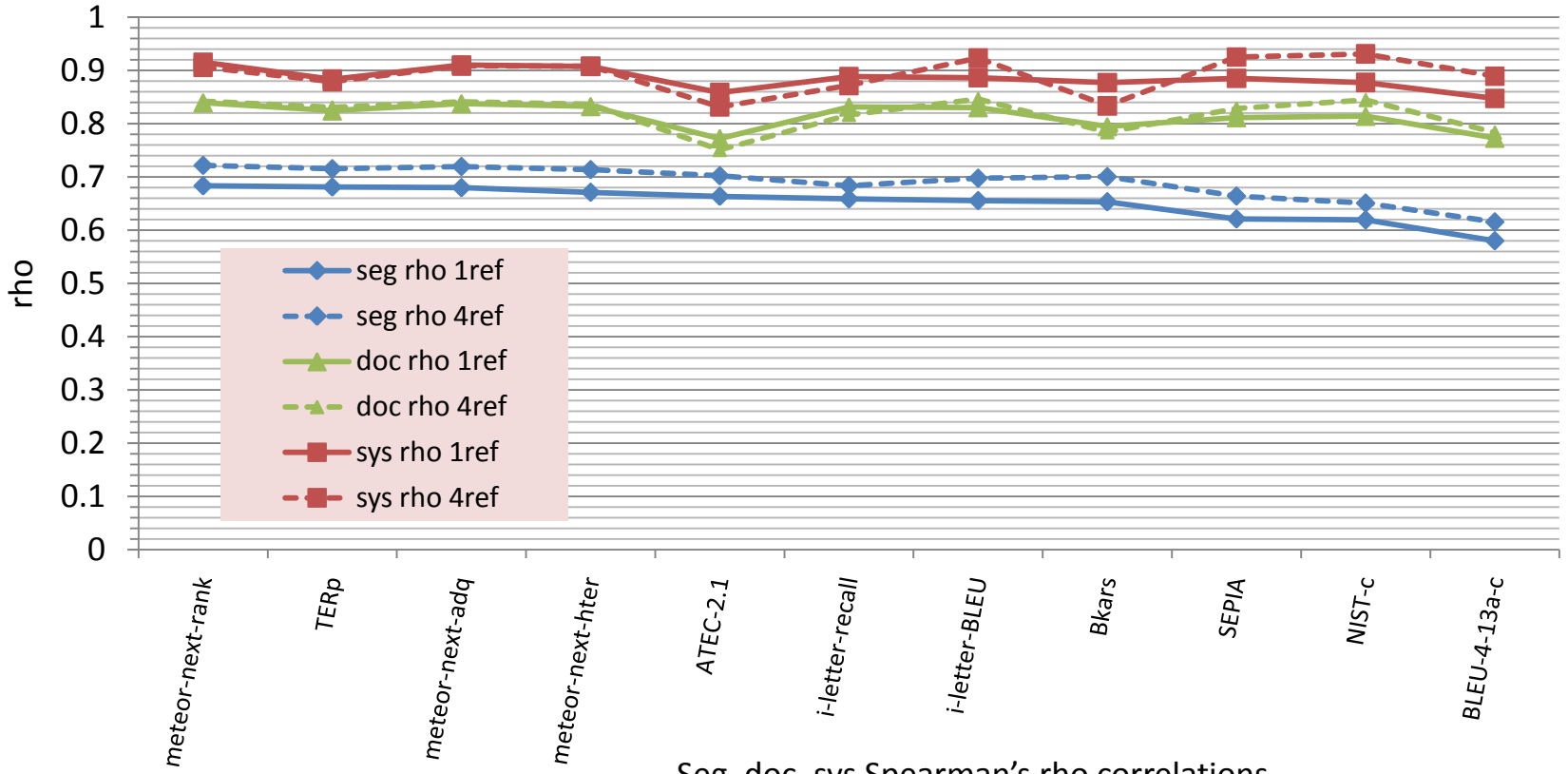
Overall Correlations

1Ref, Adequacy7, Target Eng, Sys



Overall Correlations

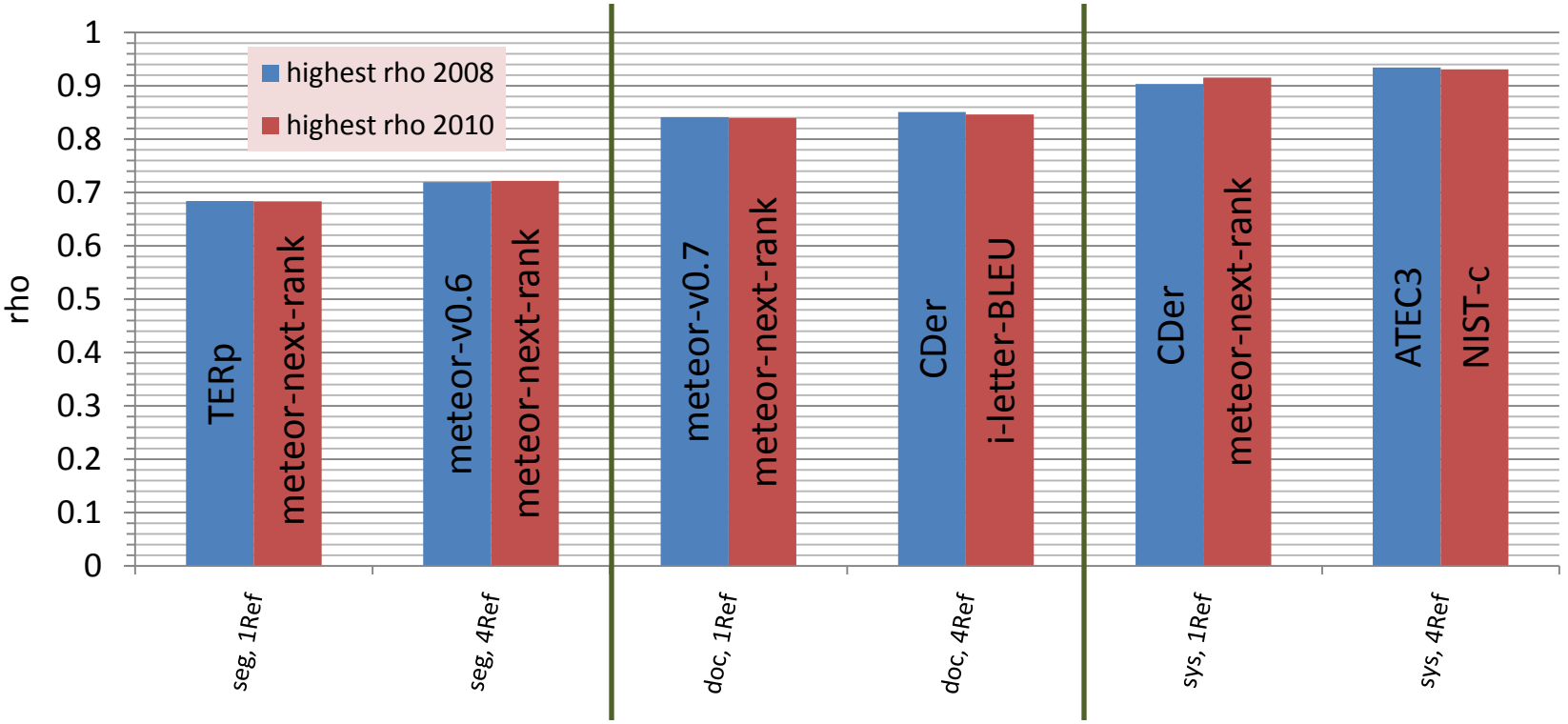
1Ref vs. 4Ref, Adequacy7, Target Eng, Seg/Doc/Sys



Seg, doc, sys Spearman's rho correlations
(absolute values)
for 11 metrics with highest 1ref seg correlation

MetricsMaTr 2008 – 2010 Highest Correlations

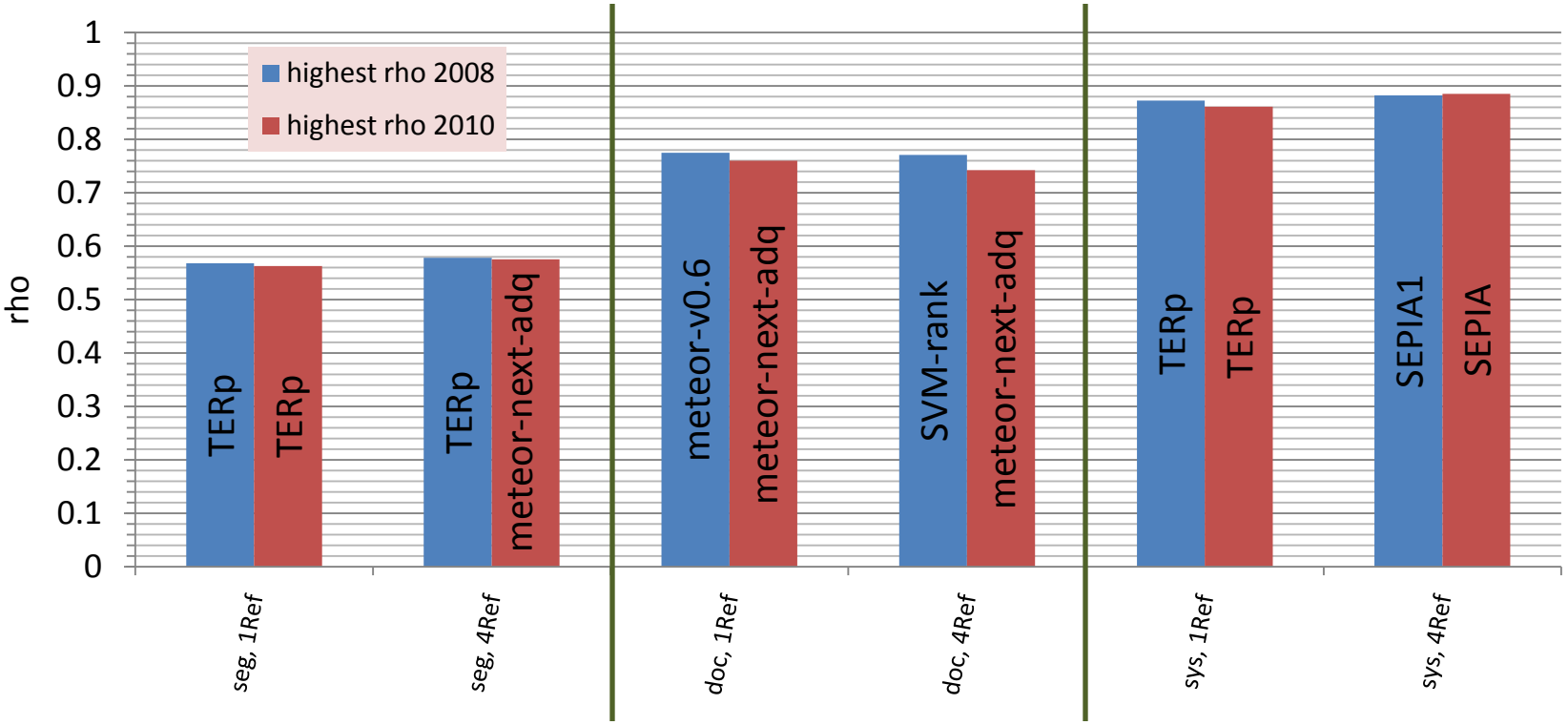
Adequacy7



Highest Spearman's rho correlations with Adequacy7 judgments

MetricsMaTr 2008 – 2010 Highest Correlations

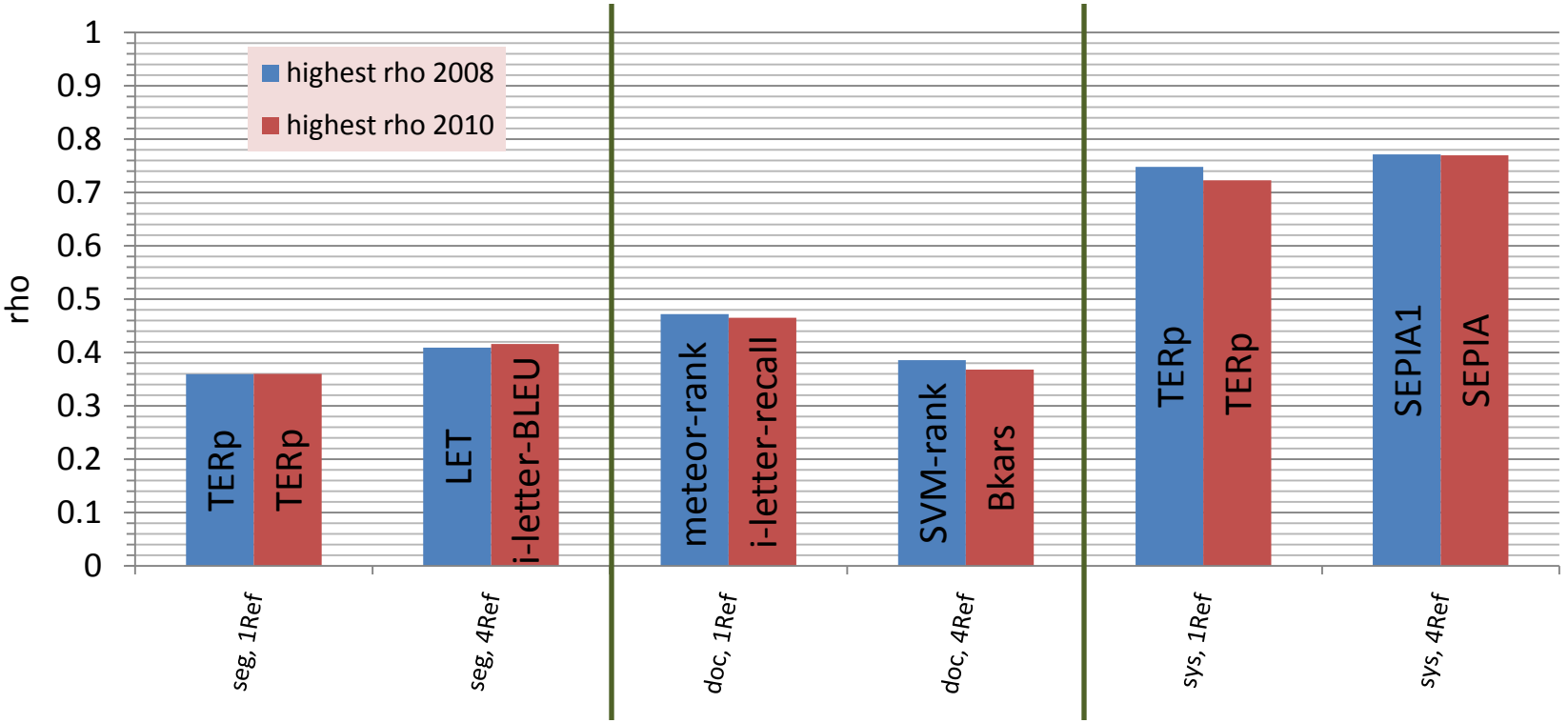
AdequacyYesNo



Highest Spearman's rho correlations with AdequacyYesNo judgments

MetricsMaTr 2008 – 2010 Highest Correlations

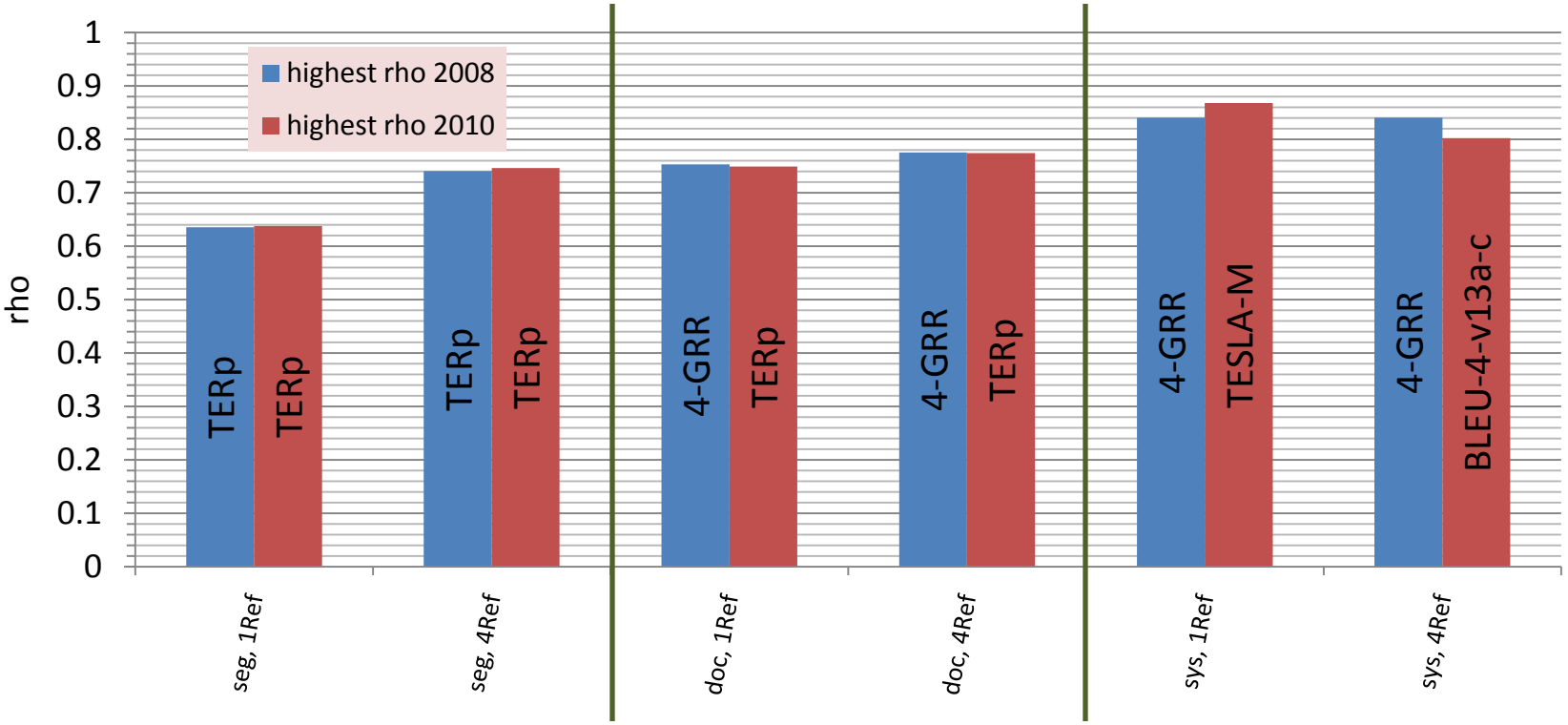
Preference



Highest Spearman's rho correlations with Preference judgments

MetricsMaTr 2008 – 2010 Highest Correlations

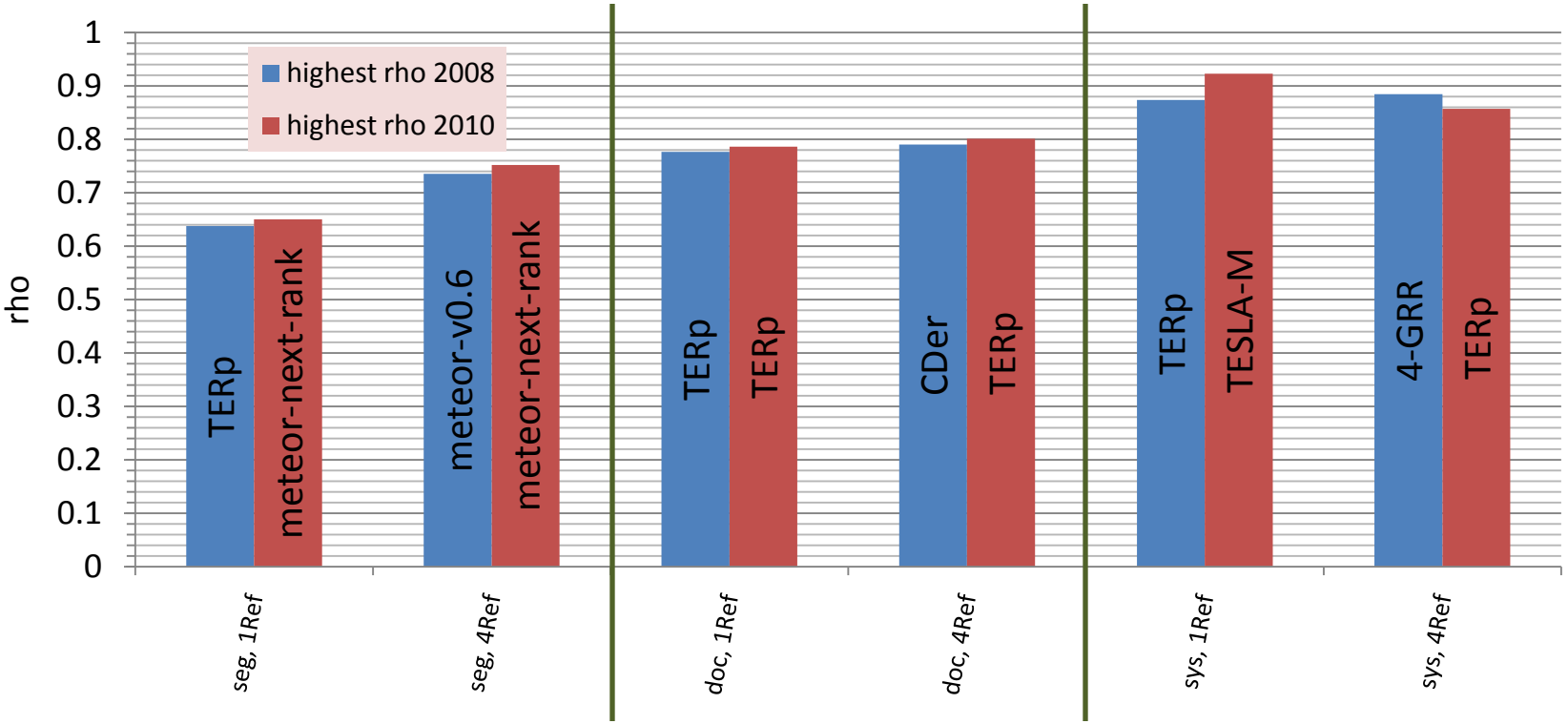
Adequacy4 (Bilingual judges – TRANSTAC Data)



Highest Spearman's rho correlations with Adequacy4 judgments

MetricsMaTr 2008 – 2010 Highest Correlations

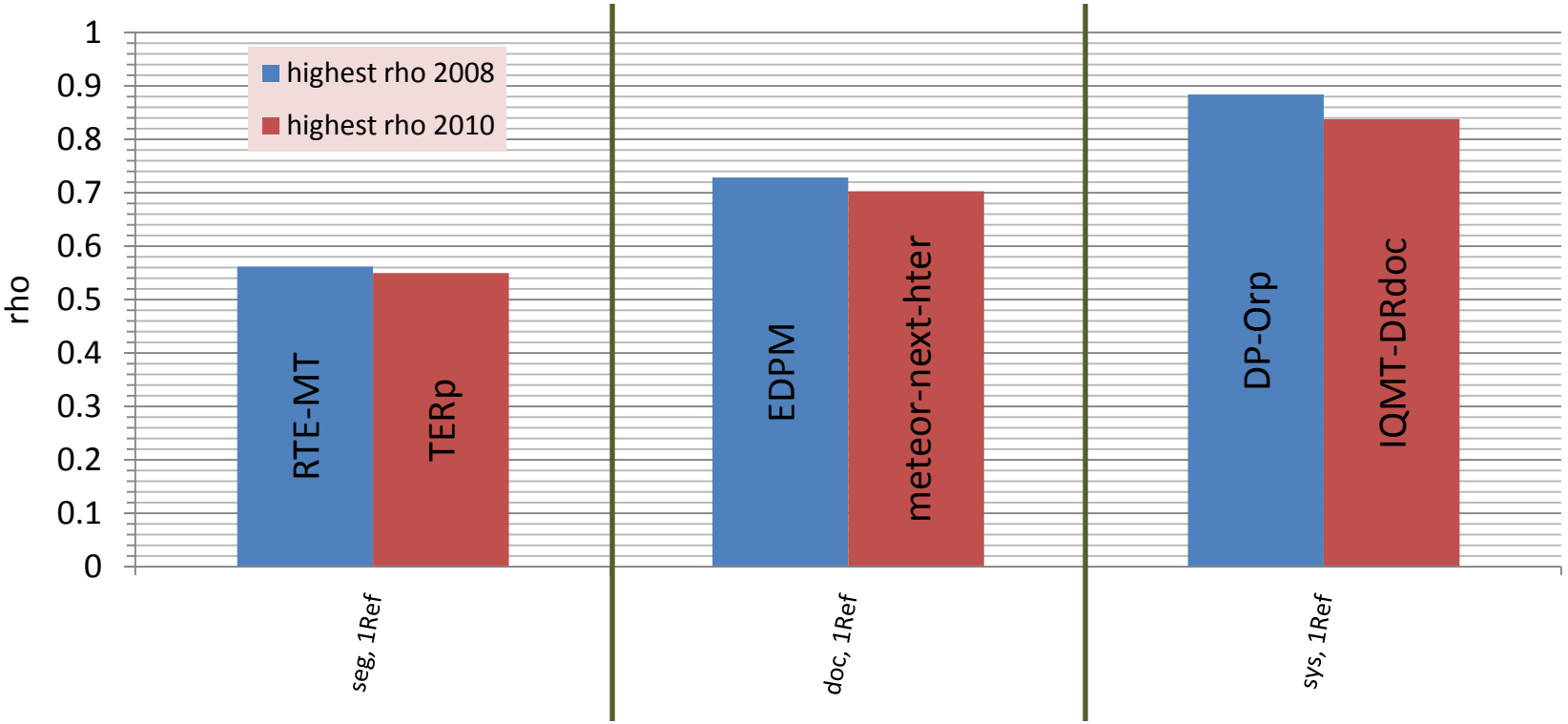
OddsConceptCorrect (Bilingual judges – TRANSTAC Data)



Highest Spearman's rho correlations with OddsConceptCorrect judgments

MetricsMaTr 2008 – 2010 Highest Correlations

HTER

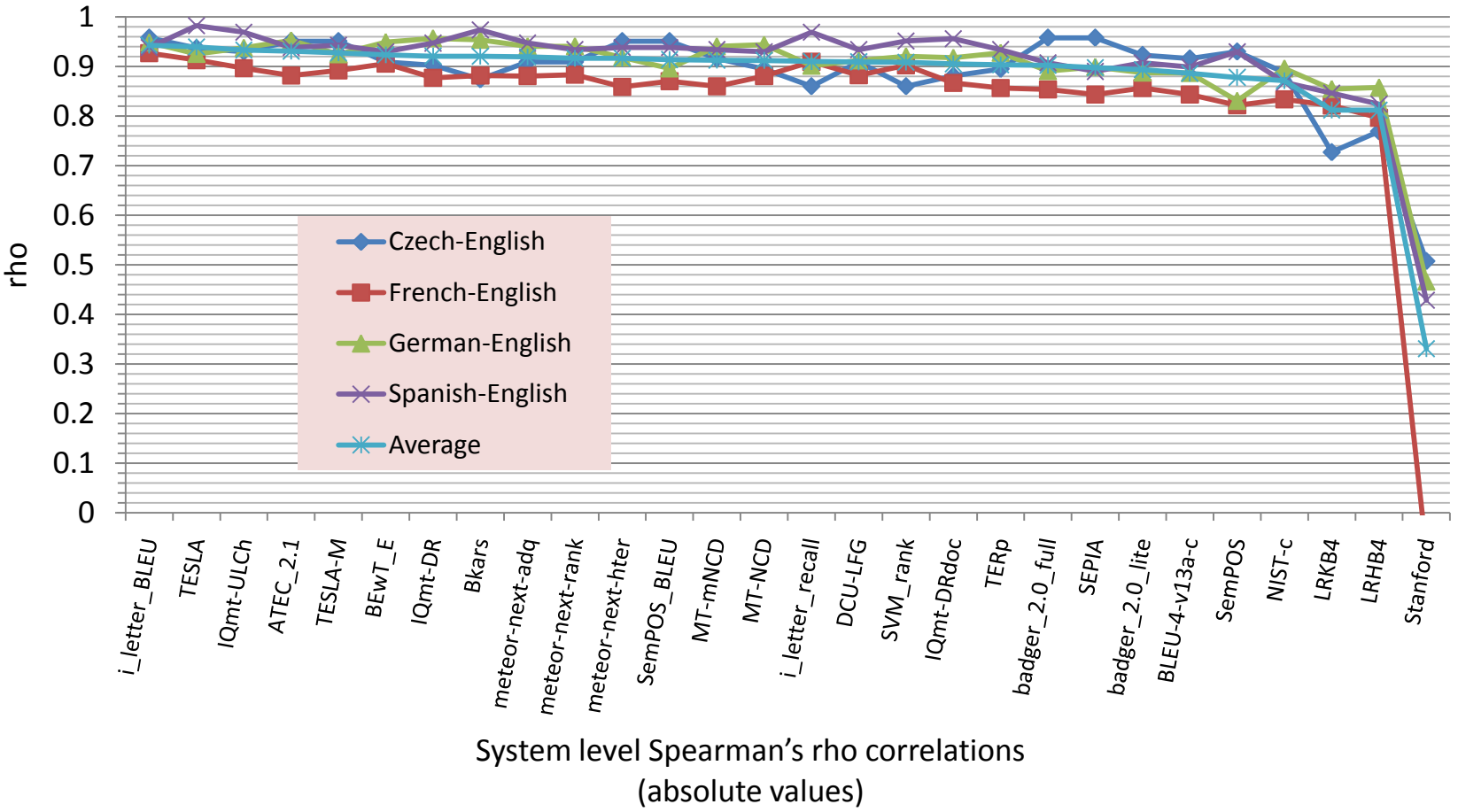


Highest Spearman's rho correlations with HTER

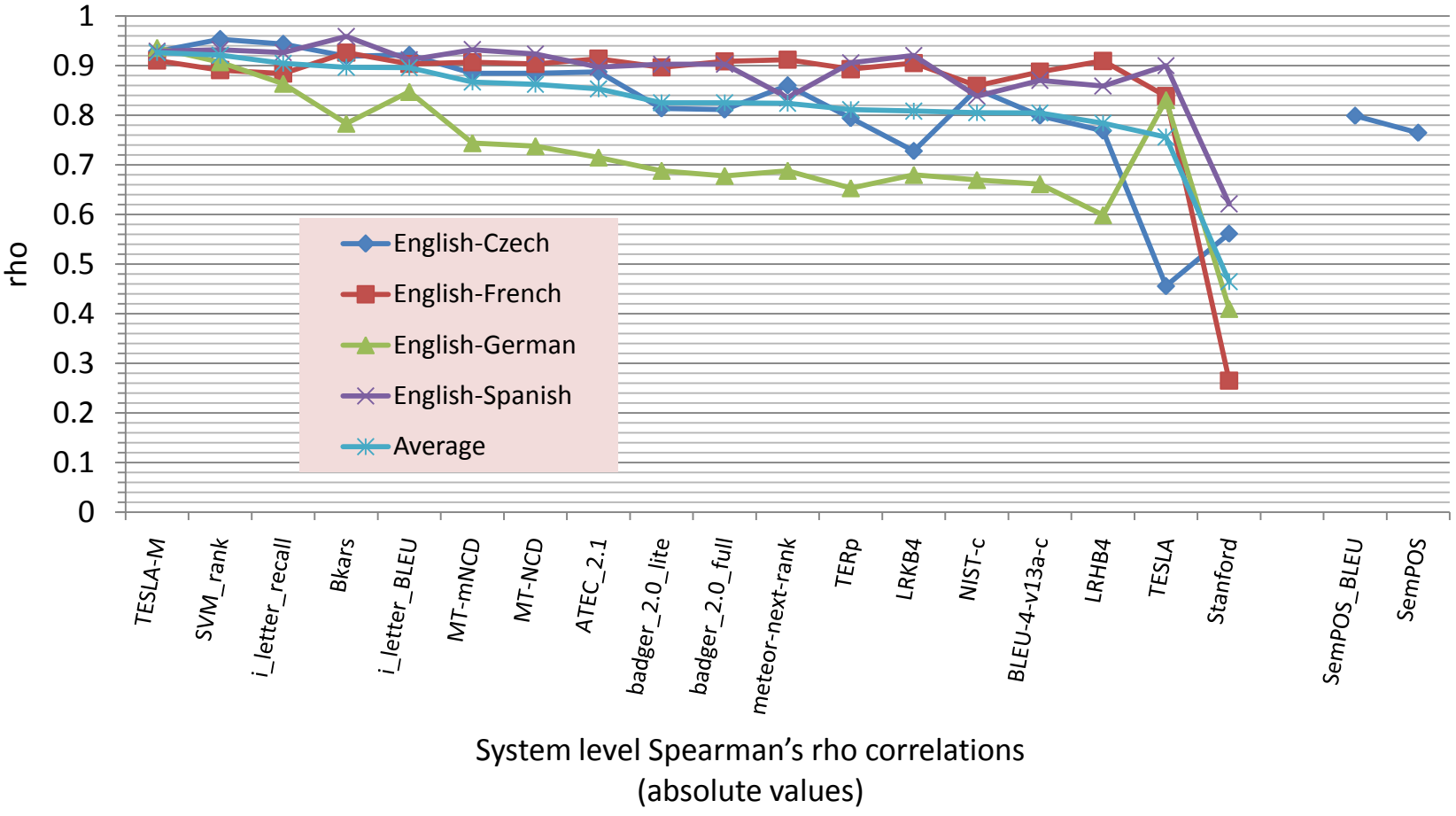
WMT10 Data Analysis

- Human assessment type: 5-system relative segment level ranking
- System level analysis:
  - System level human ranking assigned based on how many times a system's translation was judged as equal to or better than the translations of any other system
  - Correlate human ranking score with system level automatic metric scores, using Spearman's rho

WMT10 Correlations
RelativeRank, Target to-Eng, Sys



WMT10 Correlations
RelativeRank, Target from-Eng, Sys



System level Spearman's rho correlations
(absolute values)

MetricsMaTr10 Summary

- Metric approaches are somewhat converging
- Metric (upper) performance on MetricsMaTr test set similar to 2008
- More detailed data available online:
  - http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2010/results
  - http://www.statmt.org/wmt10/results