Automated Metrics for MT Evaluation

11-731:
Machine Translation
Alon Lavie
February 10, 2015

Automated Metrics for MT Evaluation

- Idea: compare output of an MT system to a "reference" good (usually human) translation: how close is the MT output to the reference translation?
- Advantages:
  - Fast and cheap, minimal human labor, no need for bilingual speakers
  - Can be used on an on-going basis during system development to test changes
  - Minimum Error-rate Training (MERT) for search-based MT approaches!
- Disadvantages:
  - Current metrics are rather crude, do not distinguish well between subtle differences in systems
  - Individual sentence scores are not very reliable, aggregate scores on a large test set are often required
- Automatic metrics for MT evaluation are an active area of current research

Similarity-based MT Evaluation Metrics

- Assess the "quality" of an MT system by comparing its output with human produced "reference" translations
- Premise: the more similar (in meaning) the translation is to the reference, the better
- Goal: an algorithm that is capable of accurately approximating this similarity
- Wide Range of metrics, mostly focusing on exact word-level correspondences:
  - Edit-distance metrics: Levenshtein, WER, PIWER, TER & HTER, others…
  - Ngram-based metrics: Precision, Recall, F1-measure, BLUE, NIST, GTM…
- Important Issue: exact word matching is very crude estimate for sentence-level similarity in meaning

Desirable Automatic Metric

- High-levels of correlation with quantified human notions of translation quality
- Sensitive to small differences in MT quality between systems and versions of systems
- Consistent – same MT system on similar texts should produce similar scores
- Reliable – MT systems that score similarly will perform similarly
- General – applicable to a wide range of domains and scenarios
- Fast and lightweight – easy to run

Automated Metrics for MT

- **Variety of Metric Uses and Applications:**
  - Compare (rank) performance of **different systems** on a common evaluation test set
  - Compare and analyze performance of different versions of **the same system**
    - Track system improvement over time
    - Which sentences got better or got worse?
  - Analyze the performance distribution of a **single system** across documents within a data set
  - Tune system parameters to optimize translation performance on a development set
- It would be nice if **one single metric** could do all of these well! But this is not an absolute necessity.
- A metric developed with one purpose in mind is likely to be used for other unintended purposes

History of Automatic Metrics for MT

- 1990s: pre-SMT, limited use of metrics from speech – WER, PI-WER…
- 2002: IBM's BLEU Metric comes out
- 2002: NIST starts MT Eval series under DARPA TIDES program, using BLEU as the official metric
- 2003: Och and Ney propose MERT for MT based on BLEU
- 2004: METEOR first comes out
- 2006: TER is released, DARPA GALE program adopts HTER as its official metric
- 2006: NIST MT Eval starts reporting METEOR, TER and NIST scores in addition to BLEU, official metric is still BLEU
- 2007: Research on metrics takes off… several new metrics come out
- 2007: MT research papers increasingly report METEOR and TER scores in addition to BLEU
- 2008: NIST and WMT introduce first comparative evaluations of automatic MT evaluation metrics
- 2009-2012: Lots of metric research… No new major winner

Automated Metric Components

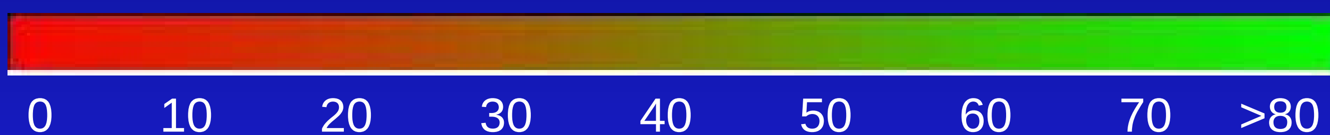
- Example:
  - Reference: "the Iraqi weapons are to be handed over to the army within two weeks"
  - MT output: "in two weeks Iraq's weapons will give army"
- Possible metric components:
  - Precision: correct words / total words in MT output
  - Recall: correct words / total words in reference
  - Combination of P and R (i.e. F1= 2PR/(P+R))
  - Levenshtein edit distance: number of insertions, deletions, substitutions required to transform MT output to the reference
- Important Issues:
  - Features: matched words, ngrams, subsequences
  - Metric: a scoring framework that uses the features
  - Perfect word matches are weak features: synonyms, inflections: "Iraq's" vs. "Iraqi", "give" vs. "handed over"

BLEU Scores - Demystified

- BLEU scores are NOT:
  - The fraction of how many sentences were translated perfectly/acceptably by the MT system
  - The average fraction of words in a segment that were translated correctly
  - Linear in terms of correlation with human measures of translation quality
  - Fully comparable across languages, or even across different benchmark sets for the same language
  - Easily interpretable by most translation professionals

BLEU Scores - Demystified

- What is TRUE about BLEU Scores:
  - Higher is Better
  - More reference human translations results in better and more accurate scores
  - General interpretability of scale:



  - Scores over 30 generally reflect understandable translations
  - Scores over 50 generally reflect good and fluent translations

The BLEU Metric

- Proposed by IBM [Papineni et al, 2002]
- Main ideas:
  - Exact matches of words
  - Match against a set of reference translations for greater variety of expressions
  - Account for Adequacy by looking at word precision
  - Account for Fluency by calculating n-gram precisions for n=1,2,3,4
  - No recall (because difficult with multiple refs)
  - To compensate for recall: introduce "Brevity Penalty"
  - Final score is weighted geometric average of the n-gram scores
  - Calculate aggregate score over a large test set
  - Not tunable to different target human measures or for different languages

The BLEU Metric

- Example:
  - Reference: "the Iraqi weapons are to be handed over to the army within two weeks"
  - MT output: "in two weeks Iraq's weapons will give army"
- BLUE metric:
  - 1-gram precision: 4/8
  - 2-gram precision: 1/7
  - 3-gram precision: 0/6
  - 4-gram precision: 0/5
  - BLEU score = 0 (weighted geometric average)

The BLEU Metric

- Clipping precision counts:
  - Reference1: “the Iraqi weapons are to be handed over to the army within two weeks”
  - Reference2: “the Iraqi weapons will be surrendered to the army in two weeks”
  - MT output: “the the the the”
  - Precision count for “the” should be “clipped” at two: max count of the word in any reference
  - Modified unigram score will be 2/4 (not 4/4)

The BLEU Metric

- Brevity Penalty:
  - Reference1: “the Iraqi weapons are to be handed over to the army within two weeks”
  - Reference2: “the Iraqi weapons will be surrendered to the army in two weeks”
  - MT output: “the Iraqi weapons will”
  - Precision score: 1-gram 4/4, 2-gram 3/3, 3-gram 2/2, 4-gram 1/1 → BLEU = 1.0
  - MT output is much too short, thus boosting precision, and BLEU doesn’t have recall...
  - An exponential Brevity Penalty reduces score, calculated based on the aggregate length (not individual sentences)

Formulae of BLEU

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

Then,

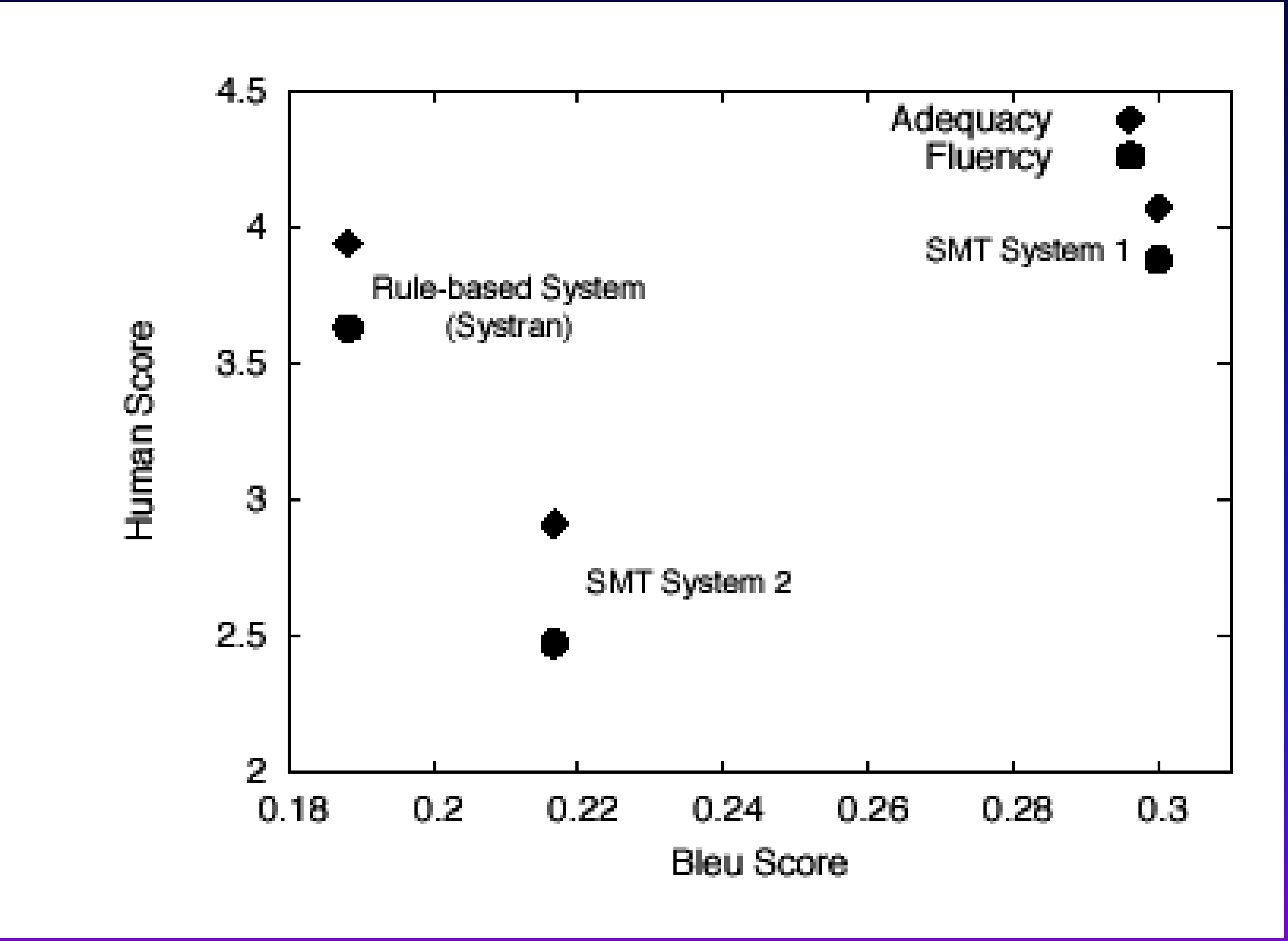
$$\mathrm{B{\scriptsize LEU}} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

$$\log \mathrm{B{\scriptsize LEU}} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n.$$

Weaknesses in BLEU

- BLUE matches word ngrams of MT-translation with multiple reference translations simultaneously → Precision-based metric
  - Is this better than matching with each reference translation separately and selecting the best match?
- BLEU Compensates for Recall by factoring in a "Brevity Penalty" (BP)
  - Is the BP adequate in compensating for lack of Recall?
- BLEU's ngram matching requires exact word matches
  - Can stemming and synonyms improve the similarity measure and improve correlation with human scores?
- All matched words weigh equally in BLEU
  - Can a scheme for weighing word contributions improve correlation with human scores?
- BLEU's higher order ngrams account for fluency and grammaticality, ngrams are geometrically averaged
  - Geometric ngram averaging is volatile to "zero" scores. Can we account for fluency/grammaticality via other means?

BLEU vs Human Scores



METEOR

- METEOR = Metric for Evaluation of Translation with Explicit Ordering [Lavie and Denkowski, 2009]
- Main ideas:
  - Combine Recall and Precision as weighted score components
  - Look only at unigram Precision and Recall
  - Align MT output with each reference individually and take score of best pairing
  - Matching takes into account translation variability via word inflection variations, synonymy and paraphrasing matches
  - Addresses fluency via a direct penalty for word order: how fragmented is the matching of the MT output with the reference?
  - Parameters of metric components **are tunable** to maximize the score correlations with human judgments for each language
- METEOR has been shown to consistently outperform BLEU in correlation with human judgments

METEOR vs BLEU

- Highlights of Main Differences:
  - METEOR word matches between translation and references includes semantic equivalents (inflections and synonyms)
  - METEOR combines *Precision and Recall* (weighted towards recall) instead of BLEU's "brevity penalty"
  - METEOR uses a direct word-ordering penalty to capture fluency instead of relying on higher order n-grams matches
  - METEOR can tune its parameters to optimize correlation with human judgments
- Outcome: METEOR has significantly better correlation with human judgments, especially at the segment-level

METEOR Components

- Unigram Precision: fraction of words in the MT that appear in the reference
- Unigram Recall: fraction of the words in the reference translation that appear in the MT
- F1= P*R/0.5*(P+R)
- Fmean = P*R/(α*P+(1-α)*R)
- Generalized Unigram matches:
  - Exact word matches, stems, synonyms, paraphrases
- Match with each reference separately and select the best match for each sentence

The Alignment Matcher

- Find the best word-to-word alignment match between two strings of words
  - Each word in a string can match at most one word in the other string
  - Matches can be based on generalized criteria: word identity, stem identity, synonymy…
  - Find the alignment of highest cardinality with minimal number of crossing branches
- Optimal search is NP-complete
  - Clever search with pruning is very fast and has near optimal results
- Earlier versions of METEOR used a greedy three-stage matching: exact, stem, synonyms
- Latest version uses an integrated single-stage search

Matcher Example

the sri lanka prime minister criticizes the leader of the country

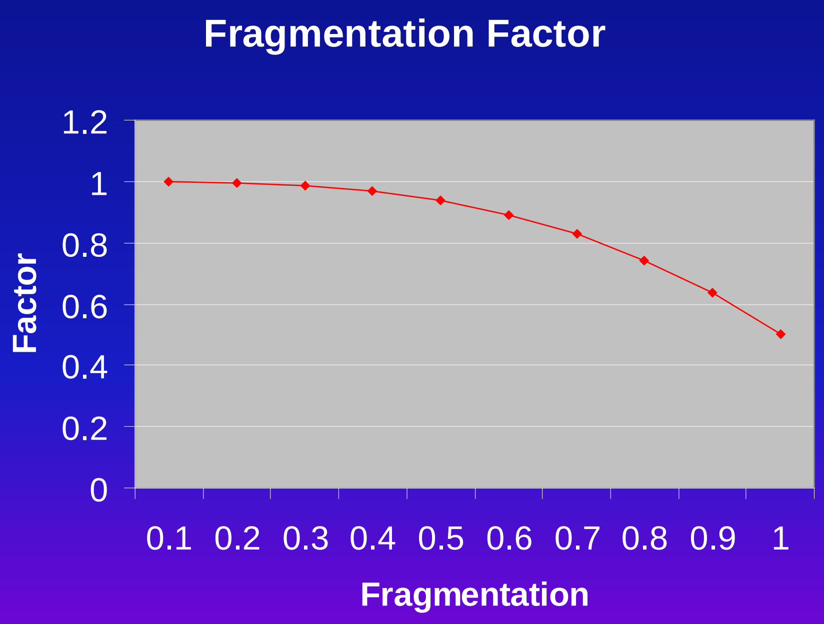
President of Sri Lanka criticized by the country’s Prime Minister

The Full METEOR Metric

- Matcher explicitly aligns matched words between MT and reference
- Matcher returns fragment count (frag) – used to calculate average fragmentation
  - (frag -1)/(length-1)
- METEOR score calculated as a discounted Fmean score
  - Discounting factor: DF = γ * (frag**β)
  - Final score: Fmean * (1- DF)
- Original Parameter Settings:
  - α= 0.9 β= 3.0 γ= 0.5
- Scores can be calculated at sentence-level
- Aggregate score calculated over entire test set (similar to BLEU)

METEOR Metric

- Effect of Discounting Factor:



METEOR Example

- Example:
  - Reference: "the Iraqi weapons are to be handed over to the army within two weeks"
  - MT output: "in two weeks Iraq's weapons will give army"
- Matching: Ref: Iraqi weapons army two weeks
  MT: two weeks Iraq's weapons army
- P = 5/8 =0.625 R = 5/14 = 0.357
- Fmean = 10*P*R/(9P+R) = 0.3731
- Fragmentation: 3 frags of 5 words = (3-1)/(5-1) = 0.50
- Discounting factor: DF = 0.5 * (frag**3) = 0.0625
- Final score:

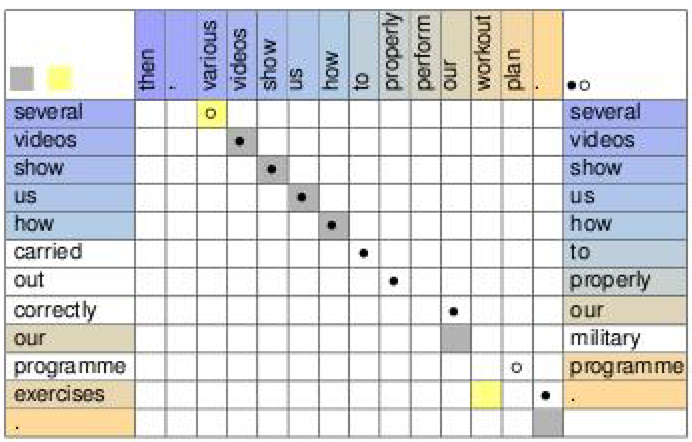
  Fmean * (1- DF) = 0.3731 * 0.9375 = 0.3498

METEOR Parameter Optimization

- METEOR has three “free” parameters that can be optimized to maximize correlation with different notions of human judgments
  - Alpha controls Precision vs. Recall balance
  - Gamma controls relative importance of correct word ordering
  - Beta controls the functional behavior of word ordering penalty score
- Optimized for Adequacy, Fluency, A+F, Rankings, and Post-Editing effort for English on available development data
- Optimized independently for different target languages
- Limited number of parameters means that optimization can be done by full exhaustive search of the parameter space

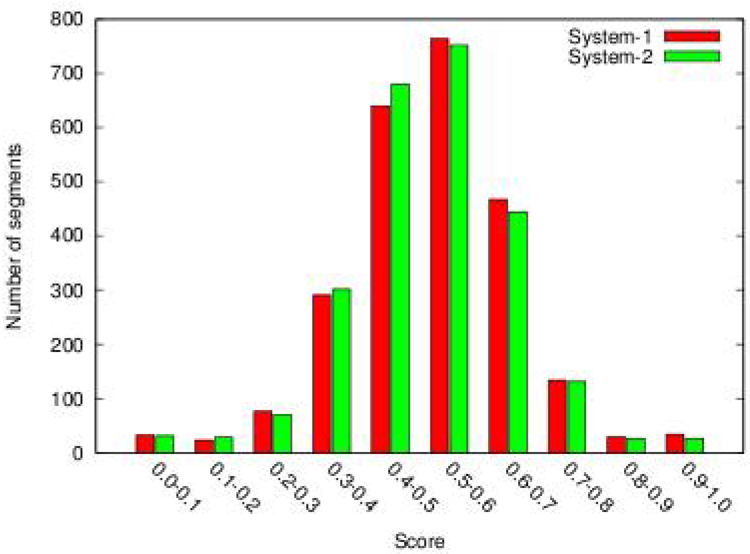
METEOR Analysis Tools

- METEOR v1.2 comes with a suite of new analysis and visualization tools called



Segment 2001

P:	0.633	vs	0.873	:	0.239
R:	0.543	vs	0.686	:	0.143
Frag:	0.231	vs	0.170	:	-0.061
Score:	0.433	vs	0.601	:	0.168



METEOR Scores - Demystified

- What is TRUE about METEOR Scores:
  - Higher is Better, scores usually higher than BLEU
  - More reference human translations help but only marginally
  - General interpretability of scale:

  - Scores over 50 generally reflect understandable translations
  - Scores over 70 generally reflect good and fluent translations

TER

- Translation Edit (Error) Rate, developed by Snover et. al. 2006
- Main Ideas:
  - Edit-based measure, similar in concept to Levenshtein distance: counts the number of word insertions, deletions and substitutions required to transform the MT output to the reference translation
  - Adds the notion of "block movements" as a single edit operation
  - Only exact word matches count, but latest version (TERp) incorporates synonymy and paraphrase matching and tunable parameters
  - Can be used as a rough post-editing measure
  - Serves as the basis for HTER – a partially automated measure that calculates TER between pre and post-edited MT output
  - Slow to run and often has a bias toward short MT translations

BLEU vs METEOR

- How do we know if a metric is better?
  - Better correlation with human judgments of MT output
  - Reduced score variability on MT outputs that are ranked equivalent by humans
  - Higher and less variable scores on scoring human translations against the reference translations

Correlation with Human Judgments

- Human judgment scores for adequacy and fluency, each [1-5] (or sum them together)
- Pearson or spearman (rank) correlations
- Correlation of metric scores with human scores at the system level
  - Can rank systems
  - Even coarse metrics can have high correlations
- Correlation of metric scores with human scores at the sentence level
  - Evaluates score correlations at a fine-grained level
  - Very large number of data points, multiple systems
  - Pearson or Spearman correlation
  - Look at metric score variability for MT sentences scored as equally good by humans

NIST Metrics MATR 2008

- First broad-scale open evaluation of automatic metrics for MT evaluation – 39 metrics submitted!!
- Evaluation period August 2008, workshop in October 2008 at AMTA-2008 conference in Hawaii
- Methodology:
  - Evaluation Plan released in early 2008
  - Data collected from various MT evaluations conducted by NIST and others
    - Includes MT system output, references and human judgments
    - Several language pairs (into English and French), data genres, and different human assessment types
  - Development data released in May 2008
  - Groups submit metrics code to NIST for evaluation in August 2008, NIST runs metrics on unseen test data
  - Detailed performance analysis done by NIST
- http://www.itl.nist.gov/iad/mig//tests/metricsmatr/2008/results/index.html

NIST Metrics MATR 2008

Origin	Source Language	Target Language	Genre(s)	Words (est.)	Systems
MT08	Arabic	English	NW, WB	15,000	10
	Chinese	English	NW, WB	15,000	10
GALE P2	Arabic	English	NW, WB	11,500	3
	Chinese	English	NW, WB	10,000	3
GALE P2.5	Arabic	English	BN	5,500	2
	Chinese	English	BC, BN	10,000	3
Transtac, Jul 07	Arabic	English	Dialog	6,500	5
	Farsi	English	Dialog	4,500	5
Transtac, Jan 07	Arabic	English	Dialog	5,000	5

NIST Metrics MATR 2008

- Human Judgment Types:
  - Adequacy, 7-point scale, straight average
  - Adequacy, Yes-No qualitative question, proportion of Yes assigned
  - Preferences, Pair-wise comparison across systems
  - Adjusted Probability that a Concept is Correct
  - Adequacy, 4-point scale
  - Adequacy, 5-point scale
  - Fluency, 5-point scale
  - HTER
- Correlations between metrics and human judgments at segment, document and system levels
- Single Reference and Multiple References
- Several different correlation statistics + confidence

NIST Metrics MATR 2008

- Human Assessment Type: **Adequacy, 7-point scale, straight average**
- Target Language: **English**
- Correlation Level: **segment**

Single Reference Track

Rank	Metric Name	Spearman's Rho		Kendall's Tau		Pearson's R	
		Value	95% confidence interval	Value	95% confidence interval	Value	95% confidence interval
1	TERp	-0.6840	(-0.6905, -0.6774)	-0.5246	(-0.5334, -0.5156)	-0.6737	(-0.6803, -0.6669)
2	METEOR-v0.6	0.6809	(0.6742, 0.6874)	0.5209	(0.5119, 0.5298)	0.6855	(0.6790, 0.6920)
3	METEOR-ranking	0.6691	(0.6622, 0.6758)	0.5132	(0.5041, 0.5222)	0.6527	(0.6456, 0.6597)
4	Meteor-v0.7	0.6652	(0.6583, 0.6720)	0.5107	(0.5016, 0.5198)	0.6789	(0.6722, 0.6855)
5	CDer	-0.6535	(-0.6605, -0.6464)	-0.4994	(-0.5086, -0.4901)	-0.6536	(-0.6606, -0.6465)
19	BLEU-4	0.5813	(0.5731, 0.5894)	0.4307	(0.4207, 0.4407)	0.5168	(0.5077, 0.5257)

NIST Metrics MATR 2008

- Human Assessment Type: **Adequacy, 7-point scale, straight average**
- Target Language: **English**
- Correlation Level: **segment**

Multiple References Track

Rank	Metric Name	Spearman's Rho		Kendall's Tau		Pearson's R	
		Value	95% confidence interval	Value	95% confidence interval	Value	95% confidence interval
1	METEOR-v0.6	0.7196	(0.7121, 0.7268)	0.5575	(0.5469, 0.5679)	0.7331	(0.7260, 0.7401)
2	SVM-Rank	0.7187	(0.7112, 0.7260)	0.5570	(0.5463, 0.5674)	0.7183	(0.7108, 0.7256)
3	Meteor-v0.7	0.7157	(0.7082, 0.7231)	0.5572	(0.5465, 0.5676)	0.7366	(0.7295, 0.7435)
4	CDer	-0.7130	(-0.7204, -0.7054)	-0.5518	(-0.5624, -0.5411)	-0.7199	(-0.7272, -0.7124)
5	TERp	-0.7127	(-0.7202, -0.7051)	-0.5488	(-0.5594, -0.5381)	-0.7216	(-0.7289, -0.7142)
19	BLEU-4	0.6203	(0.6108, 0.6297)	0.4650	(0.4529, 0.4769)	0.6064	(0.5966, 0.6159)

NIST Metrics MATR 2008

- Human Assessment Type: **Adequacy, 7-point scale, straight average**
- Target Language: **English**
- Correlation Level: **document**

Single Reference Track

Rank	Metric Name	Spearman's Rho		Kendall's Tau		Pearson's R	
		Value	95% confidence interval	Value	95% confidence interval	Value	95% confidence interval
1	Meteor-v0.7	0.8415	(0.8288, 0.8533)	0.6425	(0.6171, 0.6665)	0.8391	(0.8262, 0.8511)
2	METEOR-ranking	0.8395	(0.8267, 0.8515)	0.6403	(0.6148, 0.6644)	0.8297	(0.8162, 0.8424)
3	CDer	-0.8353	(-0.8475, -0.8221)	-0.6385	(-0.6628, -0.6130)	-0.8330	(-0.8455, -0.8197)
4	NIST-v11b	0.8143	(0.7997, 0.8280)	0.6137	(0.5868, 0.6392)	0.8096	(0.7946, 0.8236)
5	TERp	-0.8136	(-0.8273, -0.7989)	-0.6178	(-0.6432, -0.5912)	-0.8061	(-0.8203, -0.7909)
20	BLEU-4	0.7707	(0.7531, 0.7872)	0.5691	(0.5400, 0.5968)	0.7449	(0.7256, 0.7630)

NIST Metrics MATR 2008

- Human Assessment Type: **Adequacy, 7-point scale, straight average**
- Target Language: **English**
- Correlation Level: **system**

Single Reference Track

Rank	Metric Name	Spearman's Rho		Kendall's Tau		Pearson's R	
		Value	95% confidence interval	Value	95% confidence interval	Value	95% confidence interval
1	CDer	-0.9037	(-0.9359, -0.8567)	-0.7360	(-0.8187, -0.6232)	-0.8805	(-0.9201, -0.8232)
2	Meteor-v0.7	0.8968	(0.8466, 0.9311)	0.7125	(0.5920, 0.8018)	0.8745	(0.8146, 0.9159)
3	invWer	-0.8921	(-0.9280, -0.8399)	-0.7222	(-0.8088, -0.6049)	-0.8530	(-0.9012, -0.7841)
4	METEOR-ranking	0.8906	(0.8376, 0.9269)	0.7074	(0.5853, 0.7981)	0.8729	(0.8123, 0.9148)
5	TER-v0.7.25	-0.8877	(-0.9250, -0.8336)	-0.7133	(-0.8024, -0.5932)	-0.8542	(-0.9020, -0.7857)
21	BLEU-4	0.8423	(0.7689, 0.8937)	0.6512	(0.5124, 0.7568)	0.8221	(0.7407, 0.8798)

NIST Metrics MATR 2008

- Human Assessment Type: **Preferences, Pair-wise comparison across systems**
- Target Language: **English**
- Correlation Level: **segment**

Single Reference Track

Rank	Metric Name	Spearman's Rho		Kendall's Tau		Pearson's R	
		Value	95% confidence interval	Value	95% confidence interval	Value	95% confidence interval
1	TERp	-0.3597	(-0.3784, -0.3407)	-0.2569	(-0.2770, -0.2366)	-0.3403	(-0.3593, -0.3210)
2	METEOR-ranking	0.3585	(0.3394, 0.3772)	0.2550	(0.2346, 0.2751)	0.3240	(0.3045, 0.3432)
3	Meteor-v0.7	0.3551	(0.3361, 0.3739)	0.2526	(0.2322, 0.2727)	0.3409	(0.3216, 0.3599)
4	METEOR-v0.6	0.3543	(0.3352, 0.3731)	0.2520	(0.2316, 0.2721)	0.3373	(0.3180, 0.3563)
5	CDer	-0.3414	(-0.3604, -0.3222)	-0.2430	(-0.2632, -0.2225)	-0.3162	(-0.3356, -0.2966)
27	BLEU-4	0.2878	(0.2678, 0.3075)	0.2041	(0.1833, 0.2248)	0.2567	(0.2363, 0.2768)

Normalizing Human Scores

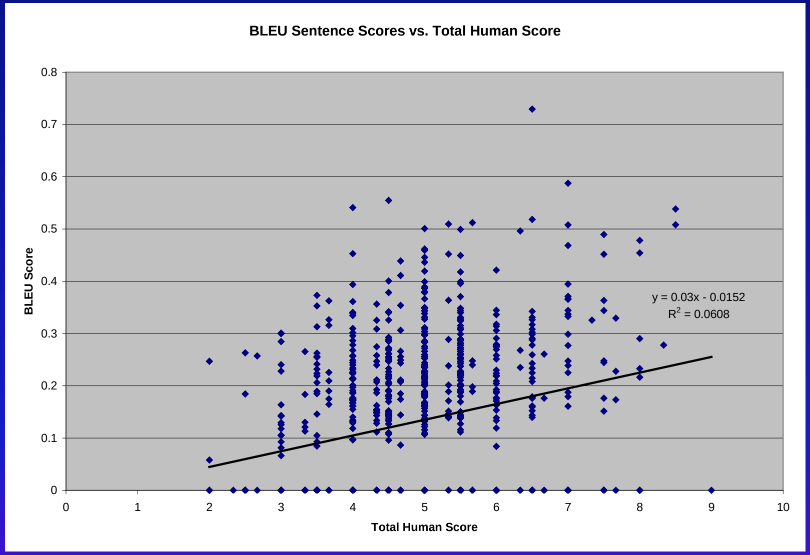
- Human scores are noisy:
  - Medium-levels of intercoder agreement, Judge biases
- MITRE group performed score normalization
  - Normalize judge median score and distributions
- Significant effect on sentence-level correlation between metrics and human scores

	Chinese data	Arabic data	Average
Raw Human Scores	0.331	0.347	0.339
Normalized Human Scores	0.365	0.403	0.384

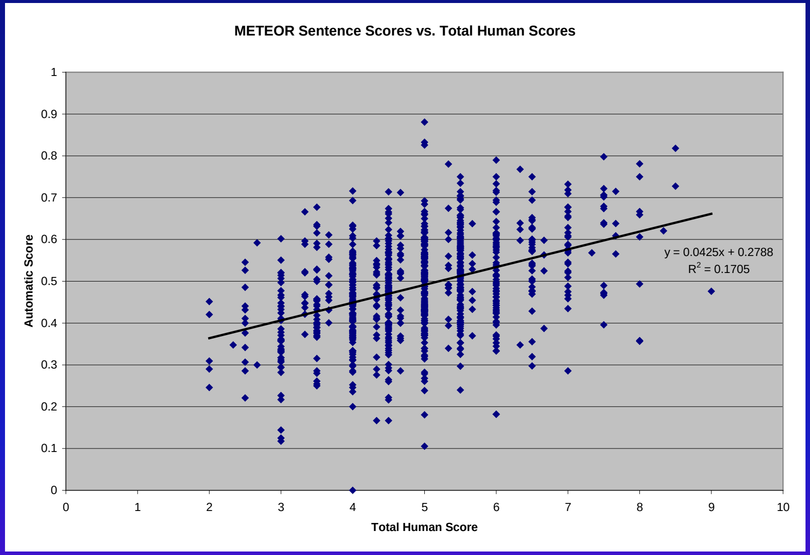
METEOR vs. BLEU Sentence-level Scores (CMU SMT System, TIDES 2003 Data)

R=0.2466

R=0.4129



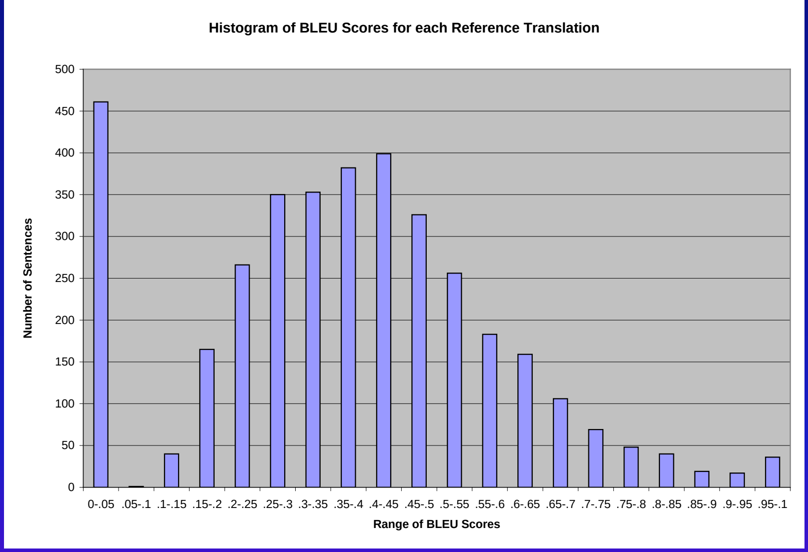
BLEU



METEOR

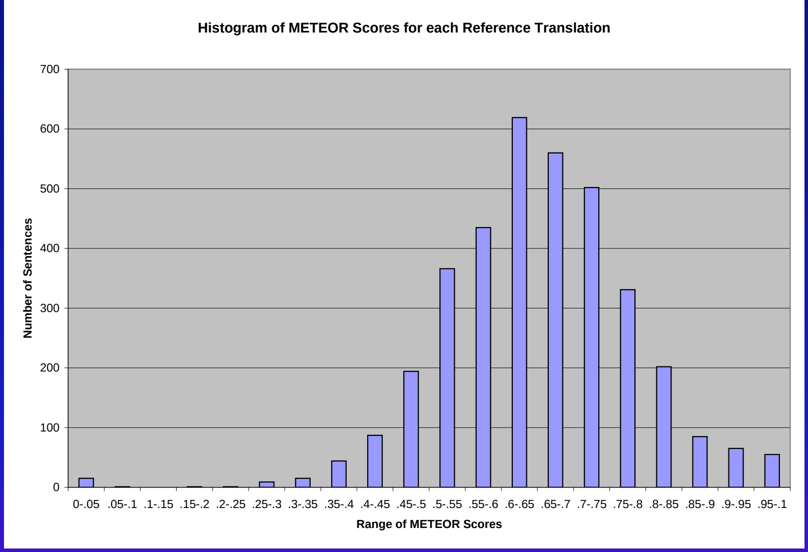
METEOR vs. BLEU
Histogram of Scores of Reference Translations
2003 Data

Mean=0.3727 STD=0.2138



BLEU

Mean=0.6504 STD=0.1310



METEOR

Testing for Statistical Significance

- MT research is experiment-driven
  - Success is measured by improvement in performance on a held-out test set compared with some baseline condition
- Methodologically important to explicitly test and validate whether any differences in aggregate test set scores are statistically significant
- One variable to control for is variance within the test data
- Typical approach: bootstrap re-sampling

Bootstrap Re-Sampling

- **Goal:** quantify impact of data distribution on the resulting test set performance score
- Establishing the true distribution of test data is difficult
- Estimated by a sampling process from the actual test set and quantifying the variance within this test set
- **Process:**
  - Sample a large number of instances from within the test set (with replacement) [e.g. 1000]
  - For each sampled test-set and condition, calculate corresponding test score
  - Repeat large number of times [e.g. 1000]
  - Calculate mean and variance
  - Establish likelihood that condition A score is better than B

Remaining Gaps

- Scores produced by most metrics are not intuitive or easy to interpret
- Scores produced at the individual segment-level are often not sufficiently reliable
- Need for greater focus on metrics with direct correlation with post-editing measures
- Need for more effective methods for mapping automatic scores to their corresponding levels of human measures (i.e. Adequacy)

Summary

- MT Evaluation is important for driving system development and the technology as a whole
- Different aspects need to be evaluated – not just translation quality of individual sentences
- Human evaluations are costly, but are most meaningful
- New automatic metrics are becoming popular, but are still rather crude, can drive system progress and rank systems
- New metrics that achieve better correlation with human judgments are being developed

References

- 2002, Papineni, K, S. Roukos, T. Ward and W-J. Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation, in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002), Philadelphia, PA, July 2002
- 2003, Och, F. J., Minimum Error Rate Training for Statistical Machine Translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003).
- 2004, Lavie, A., K. Sagae and S. Jayaraman. "The Significance of Recall in Automatic Metrics for MT Evaluation". In Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004), Washington, DC, September 2004.
- 2005, Banerjee, S. and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments" . In Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005. Pages 65-72.

References

- 2005, Lita, L. V., M. Rogati and A. Lavie, "BLANC: Learning Evaluation Metrics for MT" . In Proceedings of the Joint Conference on Human Language Technologies and Empirical Methods in Natural Language Processing (HLT/EMNLP-2005), Vancouver, Canada, October 2005. Pages 740-747.
- 2006, Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, “A Study of Translation Edit Rate with Targeted Human Annotation”. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-2006). Cambridge, MA, Pages 223–231.
- 2007, Lavie, A. and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments" . In Proceedings of the Second Workshop on Statistical Machine Translation at the 45th Meeting of the Association for Computational Linguistics (ACL-2007), Prague, Czech Republic, June 2007. Pages 228-231.
- 2008, Agarwal, A. and A. Lavie. "METEOR, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output" . In Proceedings of the Third Workshop on Statistical Machine Translation at the 46th Meeting of the Association for Computational Linguistics (ACL-2008), Columbus, OH, June 2008. Pages 115-118.

References

- 2009, Callison-Burch, C., P. Koehn, C. Monz and J. Schroeder, "*Findings of the 2009 Workshop on Statistical Machine Translation*", In Proceedings of the Fourth Workshop on Statistical Machine Translation at EACL-2009, Athens, Greece, March 2009. Pages 1-28.
- 2009, Snover, M., N. Madnani, B. Dorr and R. Schwartz, "*Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric*", In Proceedings of the Fourth Workshop on Statistical Machine Translation at EACL-2009, Athens, Greece, March 2009. Pages 259-268.

Questions?