Automated Metrics for MT Evaluation

11-731: Machine Translation
Alon Lavie
February 20, 2014
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. $F1 = \frac{2PR}{P+R}$)
  - **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 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

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

Then,

$$\text{BLEU} = BP \cdot \exp \left( \sum_{n=1}^{N} w_n \log p_n \right).$$

$$\log \text{BLEU} = \min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} w_n \log p_n.$$
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

![Graph showing BLEU vs Human Scores for different translation systems. The graph includes points for Rule-based System (Systran), SMT System 1, and SMT System 2. The x-axis represents BLEU scores, and the y-axis represents human scores. The graph compares adequacy and fluency metrics.]
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
- $F1 = \frac{P \times R}{0.5 \times (P + R)}$
- $F_{\text{mean}} = \frac{P \times R}{\alpha \times P + (1 - \alpha) \times R}$
- **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
  - \((\text{frag} - 1)/(\text{length} - 1)\)
- METEOR score calculated as a discounted Fmean score
  - Discounting factor: \(DF = \gamma \times (\text{frag}^\beta)\)
  - Final score: \(\text{Fmean} \times (1 - DF)\)
- Original Parameter Settings:
  - \(\alpha = 0.9\) \(\beta = 3.0\) \(\gamma = 0.5\)
- 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 METEOR-XRAY.
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

## NIST Metrics MATR 2008

<table>
<thead>
<tr>
<th>Origin</th>
<th>Source Language</th>
<th>Target Language</th>
<th>Genre(s)</th>
<th>Words (est.)</th>
<th>Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT08</td>
<td>Arabic</td>
<td>English</td>
<td>NW, WB</td>
<td>15,000</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>English</td>
<td>NW, WB</td>
<td>15,000</td>
<td>10</td>
</tr>
<tr>
<td>GALE P2</td>
<td>Arabic</td>
<td>English</td>
<td>NW, WB</td>
<td>11,500</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>English</td>
<td>NW, WB</td>
<td>10,000</td>
<td>3</td>
</tr>
<tr>
<td>GALE P2.5</td>
<td>Arabic</td>
<td>English</td>
<td>BN</td>
<td>5,500</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>English</td>
<td>BC, BN</td>
<td>10,000</td>
<td>3</td>
</tr>
<tr>
<td>Transl, Jul 07</td>
<td>Arabic</td>
<td>English</td>
<td>Dialog</td>
<td>6,500</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Farsi</td>
<td>English</td>
<td>Dialog</td>
<td>4,500</td>
<td>5</td>
</tr>
<tr>
<td>Transl, Jan 07</td>
<td>Arabic</td>
<td>English</td>
<td>Dialog</td>
<td>5,000</td>
<td>5</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Metric Name</th>
<th>Spearman's Rho</th>
<th>Kendall's Tau</th>
<th>Pearson's R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Value</td>
<td>95% confidence interval</td>
<td>Value</td>
</tr>
<tr>
<td>1</td>
<td>TERp</td>
<td>-0.6940</td>
<td>(-0.6905, -0.6774)</td>
<td>-0.5246</td>
</tr>
<tr>
<td>2</td>
<td>METEOR-v0.6</td>
<td>0.6909</td>
<td>(0.6742, 0.6874)</td>
<td>0.5209</td>
</tr>
<tr>
<td>3</td>
<td>METEOR-ranking</td>
<td>0.6691</td>
<td>(0.6622, 0.6758)</td>
<td>0.5132</td>
</tr>
<tr>
<td>4</td>
<td>Meteor-v0.7</td>
<td>0.6652</td>
<td>(0.6583, 0.6720)</td>
<td>0.5107</td>
</tr>
<tr>
<td>5</td>
<td>CDer</td>
<td>-0.6535</td>
<td>(-0.6605, -0.6464)</td>
<td>-0.4994</td>
</tr>
<tr>
<td>19</td>
<td>BLEU-4</td>
<td>0.5813</td>
<td>(0.5731, 0.5894)</td>
<td>0.4307</td>
</tr>
</tbody>
</table>
NIST Metrics MATR 2008

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

### Multiple References Track

<table>
<thead>
<tr>
<th>Rank</th>
<th>Metric Name</th>
<th>Spearman's Rho</th>
<th>Kendall's Tau</th>
<th>Pearson's R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Value</td>
<td>95% confidence interval</td>
<td>Value</td>
</tr>
<tr>
<td>1</td>
<td>METEOR-v0.6</td>
<td>0.7196</td>
<td>(0.7121, 0.7268)</td>
<td>0.5575</td>
</tr>
<tr>
<td>2</td>
<td>SVM-Rank</td>
<td>0.7187</td>
<td>(0.7112, 0.7260)</td>
<td>0.5570</td>
</tr>
<tr>
<td>3</td>
<td>Meteor-v0.7</td>
<td>0.7157</td>
<td>(0.7082, 0.7231)</td>
<td>0.5572</td>
</tr>
<tr>
<td>4</td>
<td>CDer</td>
<td>-0.7130</td>
<td>(-0.7204, -0.7054)</td>
<td>-0.5518</td>
</tr>
<tr>
<td>5</td>
<td>TERp</td>
<td>-0.7127</td>
<td>(-0.7202, -0.7051)</td>
<td>-0.5488</td>
</tr>
<tr>
<td>19</td>
<td>BLEU-4</td>
<td>0.6203</td>
<td>(0.6108, 0.6297)</td>
<td>0.4650</td>
</tr>
</tbody>
</table>
### NIST Metrics MATR 2008

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

#### Single Reference Track

<table>
<thead>
<tr>
<th>Rank</th>
<th>Metric Name</th>
<th>Spearman's Rho</th>
<th>Kendall's Tau</th>
<th>Pearson's R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Value</td>
<td>95% confidence interval</td>
<td>Value</td>
<td>95% confidence interval</td>
</tr>
<tr>
<td>1</td>
<td>Meteor-v0.7</td>
<td>0.8415</td>
<td>(0.8286, 0.8533)</td>
<td>0.6425</td>
</tr>
<tr>
<td>2</td>
<td>METEOR-ranking</td>
<td>0.8395</td>
<td>(0.8297, 0.8393)</td>
<td>0.6403</td>
</tr>
<tr>
<td>3</td>
<td>CDer</td>
<td>-0.6335</td>
<td>(-0.6475, -0.6221)</td>
<td>-0.6335</td>
</tr>
<tr>
<td>4</td>
<td>NIST-v11b</td>
<td>0.8143</td>
<td>(0.799, 0.820)</td>
<td>0.6137</td>
</tr>
<tr>
<td>5</td>
<td>TERp</td>
<td>-0.8136</td>
<td>(-0.8273, -0.7989)</td>
<td>-0.6178</td>
</tr>
<tr>
<td>20</td>
<td>BLEU-4</td>
<td>0.7707</td>
<td>(0.7531, 0.7872)</td>
<td>0.5691</td>
</tr>
</tbody>
</table>
NIST Metrics MATR 2008

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

### Single Reference Track

<table>
<thead>
<tr>
<th>Rank</th>
<th>Metric Name</th>
<th>Spearman's Rho</th>
<th>Kendall's Tau</th>
<th>Pearson's R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Value</td>
<td>95% confidence interval</td>
<td>Value</td>
</tr>
<tr>
<td>1</td>
<td>CDec</td>
<td>-0.9037</td>
<td>(-0.9359, -0.8567)</td>
<td>-0.7360</td>
</tr>
<tr>
<td>2</td>
<td>Meteor-0.7</td>
<td>0.8968</td>
<td>(0.8466, 0.9311)</td>
<td>0.7125</td>
</tr>
<tr>
<td>3</td>
<td>int/Wer</td>
<td>-0.8921</td>
<td>(-0.9280, -0.8399)</td>
<td>-0.7222</td>
</tr>
<tr>
<td>4</td>
<td>METEOR-ranking</td>
<td>0.8906</td>
<td>(0.8376, 0.9259)</td>
<td>0.7074</td>
</tr>
<tr>
<td>5</td>
<td>TER-0.725</td>
<td>-0.8877</td>
<td>(-0.9250, -0.8296)</td>
<td>-0.7133</td>
</tr>
<tr>
<td>21</td>
<td>BLEU-4</td>
<td>0.8423</td>
<td>(0.7689, 0.8937)</td>
<td>0.6512</td>
</tr>
</tbody>
</table>
NIST Metrics MATR 2008

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Metric Name</th>
<th>Spearman’s Rho</th>
<th>Kendall’s Tau</th>
<th>Pearson’s R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Value</td>
<td>95% confidence interval</td>
<td>Value</td>
</tr>
<tr>
<td>1</td>
<td>TERp</td>
<td>0.3597</td>
<td>(0.3534, 0.3640)</td>
<td>0.2569</td>
</tr>
<tr>
<td>2</td>
<td>METEOR-ratling</td>
<td>0.3585</td>
<td>(0.3690, 0.3772)</td>
<td>0.2550</td>
</tr>
<tr>
<td>3</td>
<td>Meteor-v0.7</td>
<td>0.3551</td>
<td>(0.3361, 0.3739)</td>
<td>0.2526</td>
</tr>
<tr>
<td>4</td>
<td>METEOR-v0.6</td>
<td>0.3543</td>
<td>(0.3352, 0.3731)</td>
<td>0.2520</td>
</tr>
<tr>
<td>5</td>
<td>CDev</td>
<td>-0.3414</td>
<td>(-0.3504, -0.3222)</td>
<td>-0.2430</td>
</tr>
<tr>
<td>27</td>
<td>BLEU-4</td>
<td>0.2678</td>
<td>(0.2678, 0.3075)</td>
<td>0.2041</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>Chinese data</th>
<th>Arabic data</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Human Scores</td>
<td>0.331</td>
<td>0.347</td>
<td>0.339</td>
</tr>
<tr>
<td>Normalized Human Scores</td>
<td>0.365</td>
<td>0.403</td>
<td>0.384</td>
</tr>
</tbody>
</table>
METEOR vs. BLEU
Sentence-level Scores
(CMU SMT System, TIDES 2003 Data)

R = 0.2466

R = 0.4129
METEOR vs. BLEU
Histogram of Scores of Reference Translations
2003 Data

Mean=0.3727 STD=0.2138

Mean=0.6504 STD=0.1310
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

References


References

Questions?