Evaluation of Machine Translation Systems: Metrics and Methodology

Alon Lavie
Language Technologies Institute
School of Computer Science
Carnegie Mellon University

IFIP Working Group 10.4
Óbidos Workshop
July 3, 2009

Outline

- Introduction: Machine Translation (MT)
- MT Evaluation: Dimensions and Approaches
- Human Evaluation Measures for MT
- Automatic Metrics for MT
  - BLEU and METEOR
- Evaluating Automatic Metrics for MT
- Challenges in Executing Meaningful MT Evaluations

Machine Translation: Where are we today?

- Age of Internet and Globalization – great demand for translation services and MT:
  - Multiple official languages of UN, EU, Canada, etc.
  - Documentation dissemination for large manufacturers (Microsoft, IBM, Apple, HP, Caterpillar, US Steel, ALCOA…)
  - Language and translation services business sector estimated at $15 Billion worldwide in 2008 and growing at a healthy pace
- Economic incentive is primarily focused on a small number of language pairs: European languages, Japanese, Chinese…
- Some increasingly decent commercial products in the market for these language pairs
  - Primarily a product of rule-based systems after many years of development
  - New generation of data-driven "statistical" MT: Google, Language Weaver
- Web-based (mostly free) MT services: Google, Babelfish, others…
- Pervasive MT between many language pairs still non-existent, but Google is trying to change that!

How Does MT Work?

- All modern MT approaches are based on building translations for complete sentences by putting together smaller pieces of translation
- Core Questions:
  - What are these smaller pieces of translation? Where do they come from?
  - How does MT put these pieces together?
  - How does the MT system pick the correct (or best) translation among many options?

Core Challenges of MT

- **Ambiguity and Language Divergences:**
  - Human languages are highly ambiguous, and differently in different languages
  - Ambiguity at all "levels": lexical, syntactic, semantic, language-specific constructions and idioms
- **Amount of required knowledge:**
  - Translation equivalencies for vast vocabularies (several 100k words and phrases)
  - Syntactic knowledge (how to map syntax of one language to another), plus more complex language divergences (semantic differences, constructions and idioms, etc.)
  - **How** do you acquire and construct a knowledge base that big that is (even mostly) correct and consistent?

Rule-based vs. Data-driven Approaches to MT

- What are the pieces of translation? Where do they come from?
  - Rule-based: large-scale "clean" word translation lexicons, manually constructed over time by experts
  - Data-driven: broad-coverage word and multi-word translation lexicons, learned automatically from available sentence-parallel corpora
- How does MT put these pieces together?
  - Rule-based: large collections of rules, manually developed over time by human experts, that map structures from the source to the target language
  - Data-driven: a computer algorithm that explores millions of possible ways of putting the small pieces together, looking for the translation that statistically looks best

Rule-based vs. Data-driven Approaches to MT

- How does the MT system pick the correct (or best) translation among many options?
  - Rule-based: Human experts encode preferences among the rules designed to prefer creation of better translations
  - Data-driven: a variety of fitness and preference scores, many of which can be learned from available training data, are used to model a total score for each of the millions of possible translation candidates; algorithm then selects and outputs the best scoring translation

Rule-based vs. Data-driven Approaches to MT

- Why have the data-driven approaches become so popular?
  - We can now do this!
    - Increasing amounts of sentence-parallel data are constantly being created on the web
    - Advances in machine learning algorithms
    - Computational power of today's computers can train systems on these massive amounts of data and can perform these massive search-based translation computations when translating new texts
  - Building and maintaining rule-based systems is too difficult, expensive and time-consuming
  - In many scenarios, it actually works better!

Statistical MT (SMT)

- Data-driven, most dominant approach in current MT research
- Proposed by IBM in early 1990s: a direct, purely statistical, model for MT
- Evolved from word-level translation to phrase-based translation
- Main Ideas:
  - Training: statistical "models" of word and phrase translation equivalence are learned automatically from bilingual parallel sentences, creating a bilingual "database" of translations
  - Decoding: new sentences are translated by a program (the decoder), which matches the source words and phrases with the database of translations, and searches the "space" of all possible translation combinations.

Statistical MT (SMT)

- Main steps in training phrase-based statistical MT:
  - Create a sentence-aligned parallel corpus
  - Word Alignment: train word-level alignment models (GIZA++)
  - Phrase Extraction: extract phrase-to-phrase translation correspondences using heuristics (Moses)
  - Minimum Error Rate Training (MERT): optimize translation system parameters on development data to achieve best translation performance
- Attractive: completely automatic, no manual rules, much reduced manual labor
- Main drawbacks:
  - Translation accuracy levels vary widely
  - Effective only with large volumes (several mega-words) of parallel text
  - Broad domain, but domain-sensitive
  - Viable only for limited number of language pairs!
- Impressive progress in last 5-10 years!

Statistical MT: Major Challenges

- Current approaches are too naïve and "direct":
  - Good at learning word-to-word and phrase-to-phrase correspondences from data
  - Not good enough at learning how to combine these pieces and reorder them properly during translation
  - Learning general rules requires much more complicated algorithms and computer processing of the data
  - The space of translations that is "searched" often doesn't contain a perfect translation
  - The fitness scores that are used aren't good enough to always assign better scores to the better translations → we don't always find the best translation even when it's there!
  - MERT is brittle, problematic and metric-dependent!
- Solutions:
  - Google solution: more and more data!
  - Research solution: "smarter" algorithms and learning methods

Rule-based vs. Data-driven MT

We thank all participants of the whole world for their comical and creative drawings; to choose the victors was not easy task!

Click here to see work of winning European of these two months, and use it to look at what the winning of USA sent us.

Rule-based

We thank all the participants from around the world for their designs cocasses and creative; selecting winners was not easy!

Click here to see the artwork of winners European of these two months, and disclosure to look at what the winners of the US have been sending.

Data-driven

Representative Example: Google Translate

- http://translate.google.com

Google Translate

Google Translate

Need for MT Evaluation

- MT Evaluation is important:
  - MT systems are becoming wide-spread, embedded in more complex systems
    - How well do they work in practice?
    - Are they reliable enough?
  - MT is a technology still in research stages
    - How can we tell if we are making progress?
    - Metrics that can drive experimental development
  - SMT's critical need for good metrics for parameter tuning (MERT)
- MT Evaluation is difficult:
  - There is no single correct translation (language variability)
  - Human evaluation is subjective
  - How good is "good enough"? Depends on application
  - Is system A better than system B? Depends on specific criteria…
- MT Evaluation is a research topic in itself! How do we assess whether an evaluation method or metric is good?

Dimensions of MT Evaluation

- Human evaluation vs. automatic metrics
- Quality assessment at sentence (segment) level vs. task-based evaluation
- "Black-box" vs. "Glass-box" evaluation
- Adequacy (is the meaning translated correctly?) vs. Fluency (is the output grammatical and fluent?) vs. Ranking (is translation-1 better than translation-2?)

Human Evaluation of MT Output

Why perform human evaluation?

- Automatic MT metrics are not sufficient:
  - What does a BLEU score of 30.0 or 50.0 mean?
  - Existing automatic metrics are crude and at times biased
  - Automatic metrics don't provide sufficient insight for error analysis
  - Different types of errors have different implications depending on the underlying task in which MT is used
- Need for reliable human measures in order to develop and assess automatic metrics for MT evaluation

Human Evaluation: Main Challenges

- Reliability and Consistency: difficulty in obtaining high-levels of intra and inter-coder agreement
  - **Intra-coder Agreement**: consistency of same human judge
  - **Inter-coder Agreement**: judgment agreement across multiple judges of quality
- Measuring Reliability and Consistency
- Developing meaningful metrics based on human judgments

Main Types of Human Assessments

- Adequacy and Fluency scores
- Human preference ranking of translations at the sentence-level
- Post-editing Measures:
  - Post-editor editing time/effort measures
  - HTER: Human Translation Edit Rate
- Human Editability measures: can humans edit the MT output into a correct translation?
- Task-based evaluations: was the performance of the MT system sufficient to perform a particular task?

Adequacy and Fluency

- Adequacy: is the **meaning** translated correctly?
  - By comparing MT translation to a reference translation (or to the source)?
- Fluency: is the output **grammatical and fluent**?
  - By comparing MT translation to a reference translation, to the source, or in isolation?
- Scales: [1-5], [1-10], [1-7], [1-4]
- Initiated during DARPA MT evaluations during mid-1990s
- Most commonly used until recently
- Main Issues: definitions of scales, agreement, normalization across judges

Human Preference Ranking of MT Output

- Method: compare two or more translations of the same sentence and rank them in quality
  - More intuitive, less need to define exact criteria
  - Can be problematic: comparing bad long translations is very confusing and unreliable
- Main Issues:
  - Binary rankings or multiple translations?
  - Agreement levels
  - How to use ranking scores to assess systems?

Human Assessment in WMT-09

- WMT-09: Shared task on developing MT systems between several European languages (to English and from English)
- Also included a system combination track and an automatic MT metric evaluation track
- Official Metric: Human Preference Rankings
- Detailed evaluation and analysis of results
- 2-day Workshop at EACL-09, including detailed analysis paper by organizers

Human Rankings at WMT-09

- **Instructions:** Rank translations from Best to Worst relative to the other choices (ties are allowed)
- Annotators were shown at most five translations at a time.
- For most language pairs there were more than 5 systems submissions. No attempt to get a complete ordering over all the systems at once
- Relied on random selection and a reasonably large sample size to make the comparisons fair.
- **Metric to compare MT systems**: Individual systems and system combinations are ranked based on how frequently they were judged to be better than or equal to any other system.

Assessing Coding Agreement

- **Intra**-annotator Agreement:
  - 10% of the items were repeated and evaluated twice by each judge.
- **Inter**-annotator Agreement:
  - 40% of the items were randomly drawn from a common pool that was shared across all annotators creating a set of items that were judged by multiple annotators.
- Agreement Measure: Kappa Coefficient

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

P(A) is the proportion of times that the annotators agree
P(E) is the proportion of time that they would agree by chance.

Assessing Coding Agreement

INTER-ANNOTATOR AGREEMENT

Evaluation type	$P(A)$	$P(E)$	K
Sentence ranking	.549	.333	.323
Yes/no to edited output	.774	.5	.549

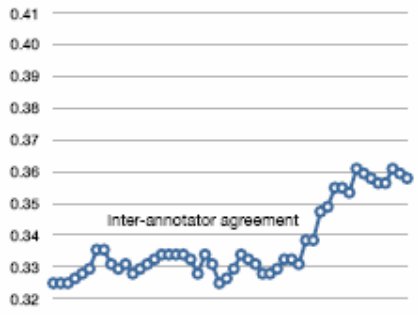
INTRA-ANNOTATOR AGREEMENT

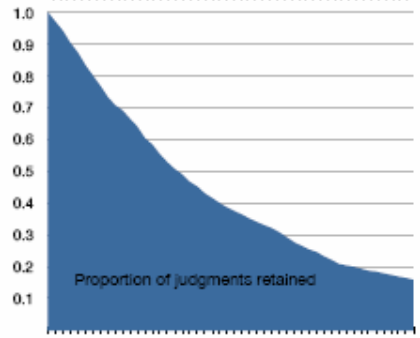
Evaluation type	$P(A)$	$P(E)$	K
Sentence ranking	.707	.333	.561
Yes/no to edited output	.866	.5	.732

Table 4: Inter- and intra-annotator agreement for the two types of manual evaluation

Common Interpretation of Kappa Values:

0.0-0.2: slight agreement
0.2-0.4: fair agreement
0.4-0.6: moderate agreement
0.6-0.8: substantial agreement
0.8-1.0: near perfect agreement





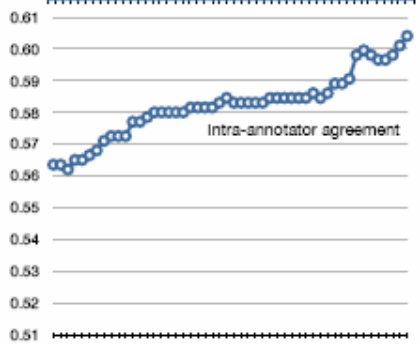
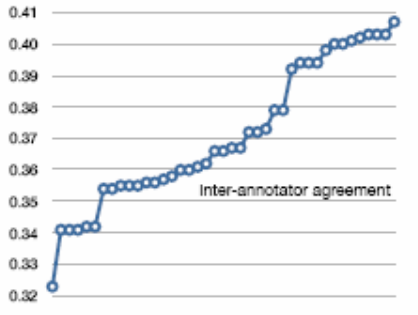
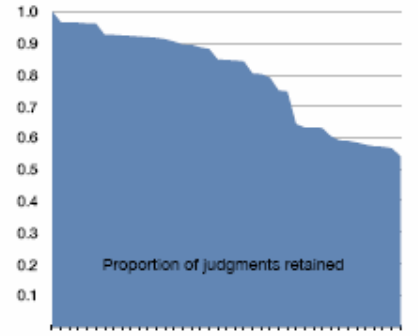


Figure 4: The effect of discarding every annotators' initial judgments, up to the first 50 items





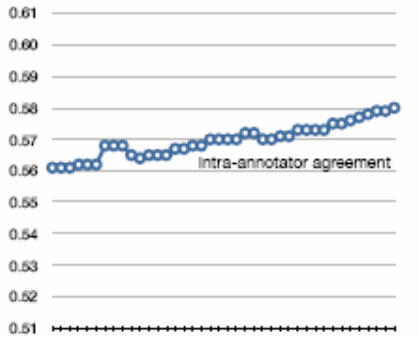


Figure 5: The effect of removing annotators with the lowest agreement, disregarding up to 40 annotators

Assessing MT Systems

- Human Rankings were used to assess:
  - Which systems produced the best translation quality for each language pair?
  - Did the system combinations produce better translations than individual systems?
  - Which of the systems that used only the provided training materials produced the best translation quality?

French–English
980 pairwise judgments per system

System	C?	≥others
GOOGLE ●	no	.76
DCU ★	yes	.66
LIMSI ●	no	.65
JHU ★	yes	.62
UEDIN ★	yes	.61
UKA	yes	.61
LIUM-SYSTRAN	no	.60
RBMT5	no	.59
CMU-STATXFER ★	yes	.58
RBMT1	no	.56
USAAR	no	.55
RBMT3	no	.54
RWTH ★	yes	.52
COLUMBIA	yes	.50
RBMT4	no	.47
GENEVA	no	.34

English-French
564 pairwise judgments per system

System	C?	≥others
LIUM-SYSTRAN ●	no	.73
GOOGLE ●	no	.68
UKA ●★	yes	.66
SYSTRAN ●	no	.65
RBMT3 ●	no	.65
DCU ●★	yes	.65
LIMSI ●	no	.64
UEDIN ★	yes	.60
RBMT4	no	.59
RWTH	yes	.58
RBMT5	no	.57
RBMT1	no	.54
USAAR	no	.48
GENEVA	no	.38

German–English
936 pairwise judgments per system

System	C?	≥others
RBMT5	no	.66
USAAR ●	no	.65
GOOGLE ●	no	.65
RBMT2 ●	no	.64
RBMT3	no	.64
RBMT4	no	.62
STUTTGART ●★	yes	.61
SYSTRAN ●	no	.60
UEDIN ★	yes	.59
UKA ★	yes	.58
UMD ★	yes	.56
RBMT1	no	.54
LIU ★	yes	.50
RWTH	yes	.50
GENEVA	no	.33
JHU-TROMBLE	yes	.13

English–German
1232 pairwise judgments per system

System	C?	≥others
RBMT2 ●	no	.66
RBMT3 ●	no	.64
RBMT5 ●	no	.64
USAAR	no	.58
RBMT4	no	.58
RBMT1	no	.57
GOOGLE	no	.54
UKA ★	yes	.54
UEDIN ★	yes	.51
LIU ★	yes	.49
RWTH ★	yes	.48
STUTTGART	yes	.43

Automatic 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 very 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 very 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

Automatic Metrics for MT Evaluation

- 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"

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
- Not "Game-able" – not easily susceptible to manipulation and cheating
- Fast and lightweight – easy to run

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

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

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,

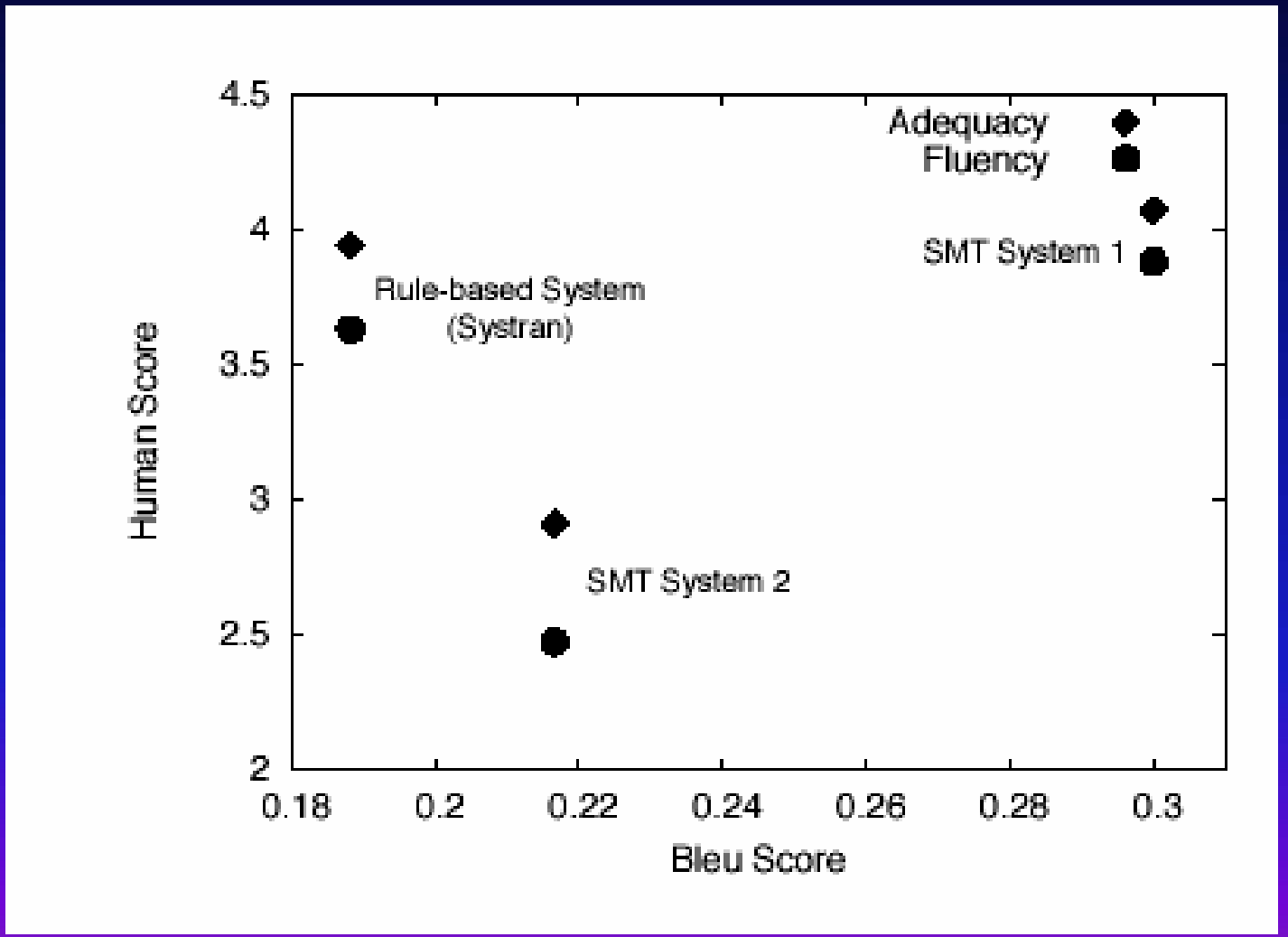
$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

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

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



The METEOR Metric

- Metric developed by Lavie et al. at CMU/LTI: METEOR = Metric for Evaluation of Translation with Explicit Ordering
- Main new ideas:
  - Include both Recall and Precision as 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 word variatiability (via stemming) and synonyms
  - Address fluency via a direct scoring component: matching fragmentation
  - Tuning of scoring component weights to optimize 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

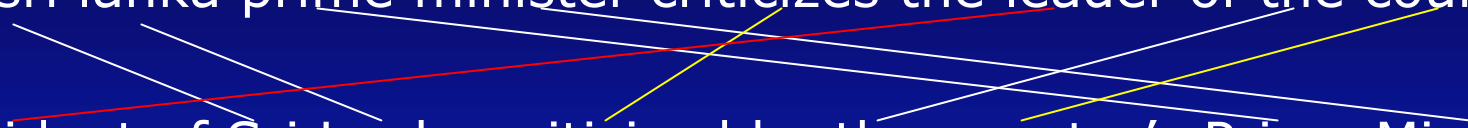
- Unigram Precision: fraction of words in the MT that appear in the reference
- Unigram Recall: fraction of the words in the reference translation that appear in the MT
- F1= P*R/0.5*(P+R)
- Fmean = P*R/(α*P+(1-α)*R)
- Generalized Unigram matches:
  - Exact word matches, stems, synonyms
- Match with 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
- Greedy three-stage matching: exact, stem, synonyms

Matcher Example

the sri lanka prime minister criticizes the leader of the country



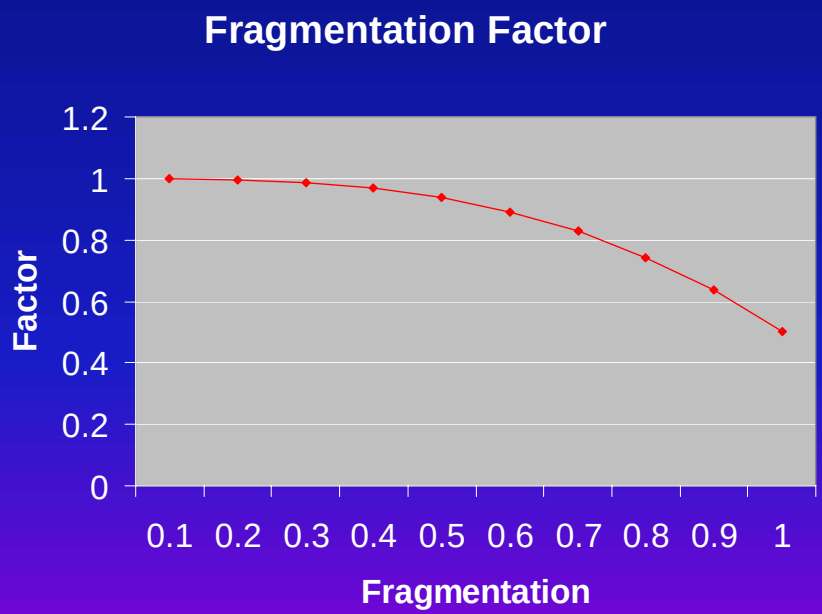
President of Sri Lanka criticized by the country’s Prime Minister

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

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

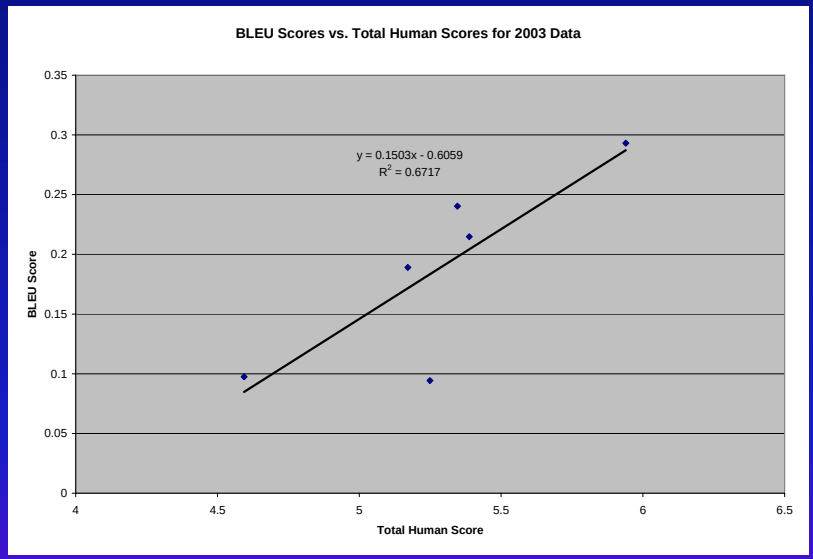
- Human judgment scores for adequacy and fluency, each [1-5] (or sum them together)
- Pearson or spearman (rank) correlations
- Correlation of metric scores with human scores at the system level
  - Can rank systems
  - Even coarse metrics can have high correlations
- Correlation of metric scores with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

Evaluation Setup

- Data: LDC Released Common data-set (DARPA/TIDES 2003 Chinese-to-English and Arabic-to-English MT evaluation data)
- Chinese data:
  - 920 sentences, 4 reference translations
  - 7 systems
- Arabic data:
  - 664 sentences, 4 reference translations
  - 6 systems
- Metrics Compared: BLEU, P, R, F1, Fmean, METEOR (with several features)

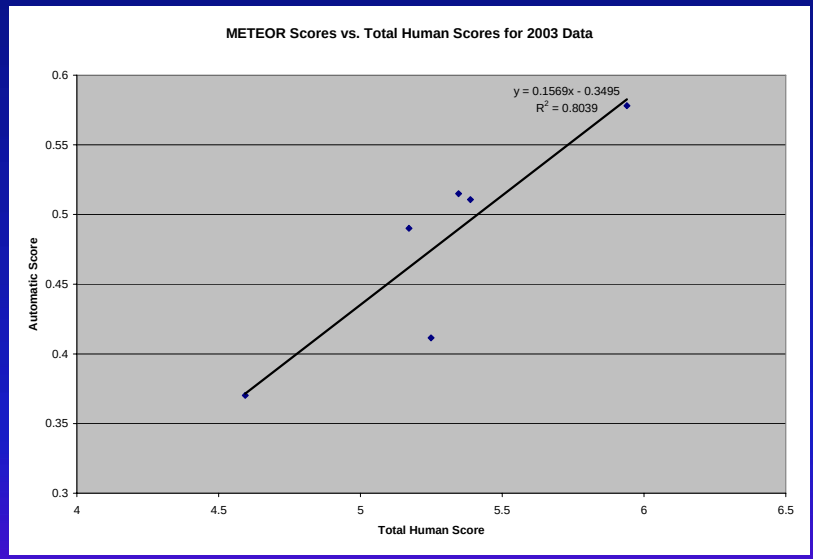
METEOR vs. BLEU: 2003 Data, System Scores

R=0.8196



BLEU

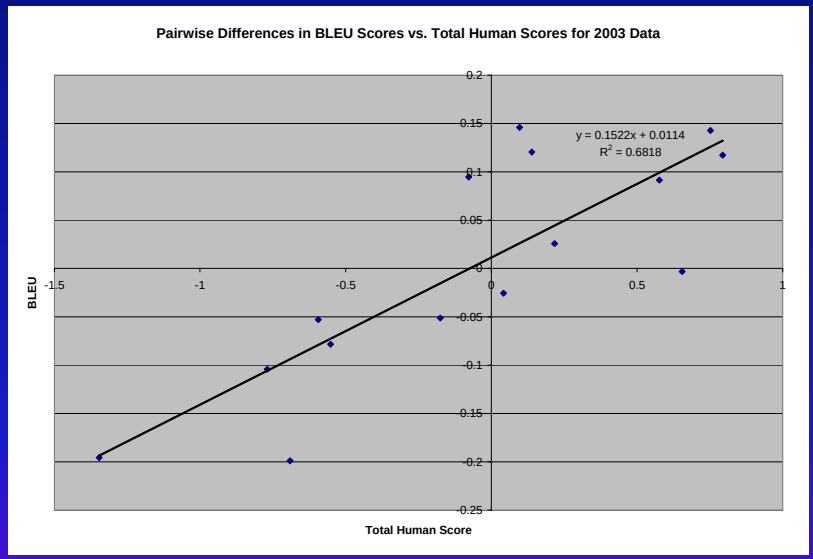
R=0.8966



METEOR

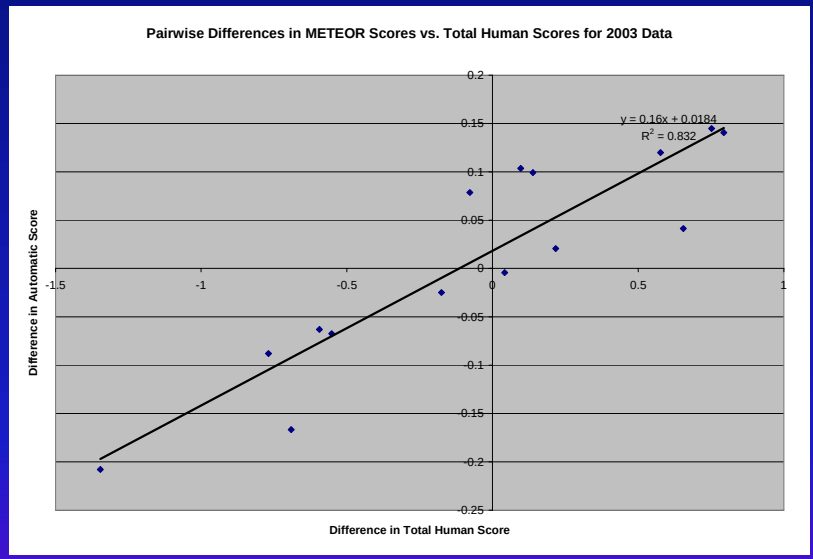
METEOR vs. BLEU: 2003 Data, Pairwise System Scores

R=0.8257



BLEU

R=0.9121



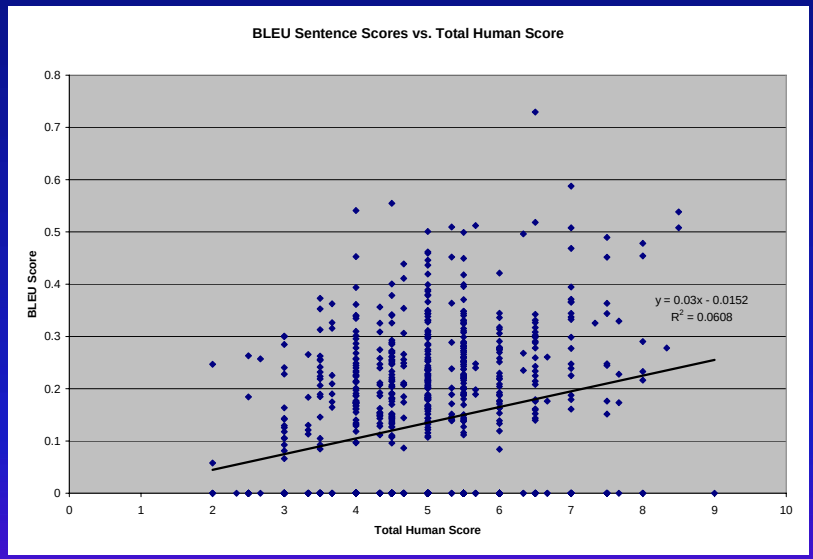
METEOR

Evaluation Results: System-level Correlations

	Chinese data	Arabic data	Average
BLEU	0.828	0.930	0.879
Mod-BLEU	0.821	0.926	0.874
Precision	0.788	0.906	0.847
Recall	0.878	0.954	0.916
F1	0.881	0.971	0.926
Fmean	0.881	0.964	0.922
METEOR	0.896	0.971	0.934

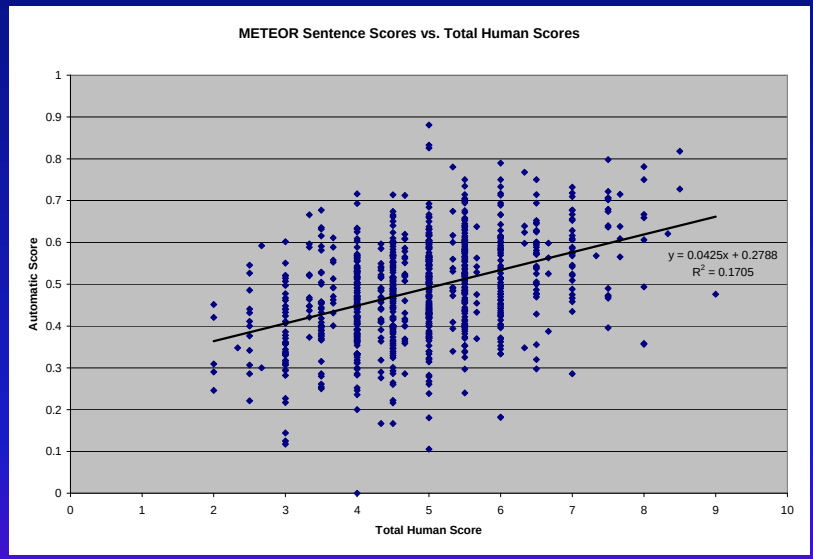
METEOR vs. BLEU Sentence-level Scores (CMU SMT System, TIDES 2003 Data)

R=0.2466



BLEU

R=0.4129



METEOR

Evaluation Results: Sentence-level Correlations

	Chinese data	Arabic data	Average
BLEU	0.194	0.228	0.211
Mod-BLEU	0.285	0.307	0.296
Precision	0.286	0.288	0.287
Recall	0.320	0.335	0.328
Fmean	0.327	0.340	0.334
METEOR	0.331	0.347	0.339

Adequacy, Fluency and Combined: Sentence-level Correlations Arabic Data

	Adequacy	Fluency	Combined
BLEU	0.239	0.171	0.228
Mod-BLEU	0.315	0.238	0.307
Precision	0.306	0.210	0.288
Recall	0.362	0.236	0.335
Fmean	0.367	0.240	0.340
METEOR	0.370	0.252	0.347

METEOR Mapping Modules: Sentence-level Correlations

	Chinese data	Arabic data	Average
Exact	0.293	0.312	0.303
Exact+Pstem	0.318	0.329	0.324
Exact+WNste	0.312	0.330	0.321
Exact+Pstem +WNsyn	0.331	0.347	0.339

Normalizing Human Scores

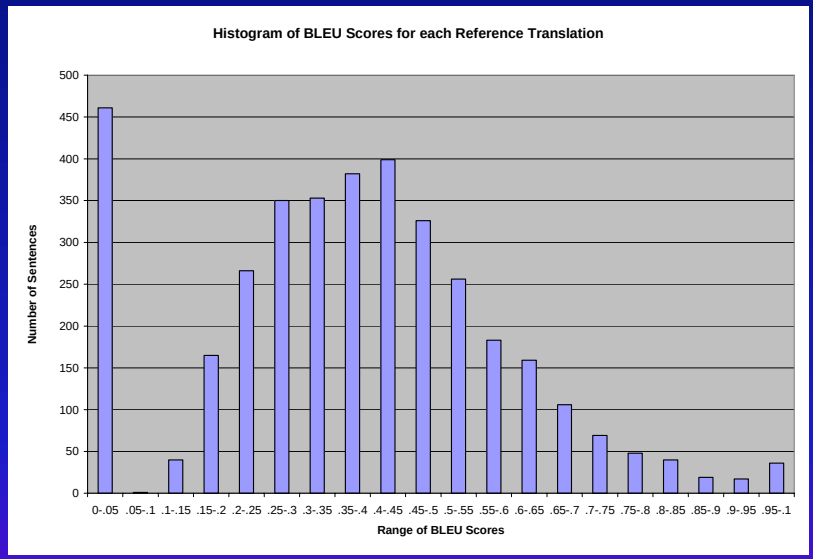
- Human scores are noisy:
  - Medium-levels of intercoder agreement, Judge biases
- MITRE group performed score normalization
  - Normalize judge median score and distributions
- Significant effect on sentence-level correlation between metrics and human scores

	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

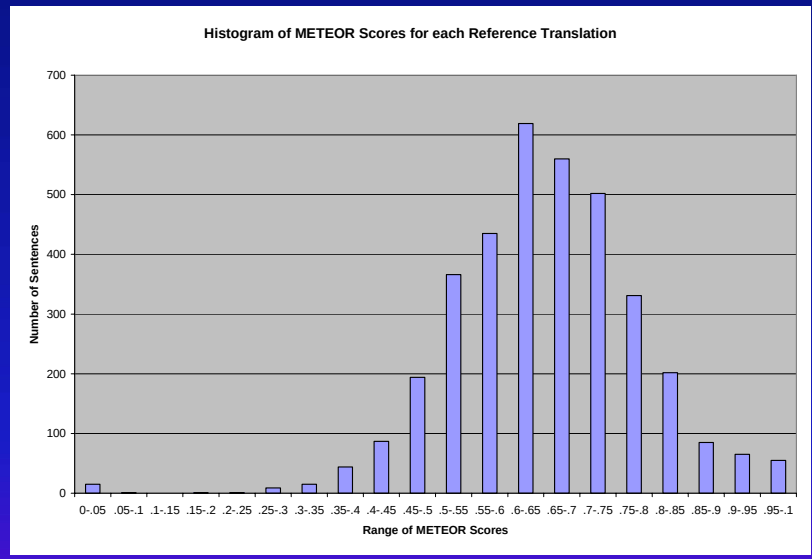
Histogram of Scores of Reference Translations
2003 Data

Mean=0.3727 STD=0.2138



BLEU

Mean=0.6504 STD=0.1310



METEOR

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, and Rankings for English on available development data
- Optimized for languages other than English
- Limited number of parameters means that optimization can be done by full exhaustive search of the parameter space

METEOR Parameter Optimization

- Optimizing for Adequacy, Fluency and A+F
- Original Parameters:
  - α= 0.9 β= 3.0 γ= 0.5

Table II. Optimal Values of Tuned Parameters for Different Criteria in English

	Adequacy	Fluency	Sum
α	0.82	0.78	0.81
β	1.0	0.75	0.83
γ	0.21	0.38	0.28

METEOR Parameter Optimization

- Optimizing for other languages

Table III. Optimal Values of Tuned Parameters for Different Criteria Across Languages

	Adequacy	Fluency	Sum
French: α	0.86	0.74	0.76
β	0.5	0.5	0.5
γ	1.0	1.0	1.0
German: α	0.95	0.95	0.95
β	0.5	0.5	0.5
γ	0.6	0.8	0.75
Spanish: α	0.95	0.62	0.95
β	1.0	1.0	1.0
γ	0.9	1.0	0.98

METEOR Parameter Optimization

- Optimizing for Ranking

Table V. Optimal Values of Tuned Parameters for Ranking

	English	German	French	Spanish
α	0.95	0.9	0.9	0.9
β	0.5	3	0.5	0.5
γ	0.45	0.15	0.55	0.55

Table VI. Average Spearman Correlation with Human Rankings for METEOR on Development Data

	Original	Re-tuned
English	0.3813	0.4020
German	0.2166	0.2838
French	0.2992	0.3640
Spanish	0.2021	0.2186

NIST Metrics MATR

- 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

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

- 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

- 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

- 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

- 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

- 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

- 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)

Comparative MT Evaluations: Challenges and Pitfalls

- Test set changes from year to year:
  - Unlike metrics, MT systems are too complex to be "submitted" for evaluation by a central party
  - To guard against cheating, a new test set is created every year
  - Significant effort to try ensure comparable test set difficulty levels
  - Results comparable across systems, but difficult to assess year-to-year progress
- NIST MT Eval solution:
  - Two test sets "comparative" and "progress"
  - Only trusted parties allowed to participate on "progress" set, must delete all traces of data after the evaluation
- DARPA GALE solution:
  - Try hard to control for test set difficulty
  - Reuse some "sequestered" material from previous years evaluation to select new test sets of similar difficulty

Comparative MT Evaluations: Challenges and Pitfalls

- What is the "official" metric for the evaluation?
  - Should the official metric be a human judgment type? Which one?
  - Should there even be **one single** official metric? Why not run several?
  - Is it important that systems be able to tune to the official evaluation metric? Does that introduce a bias against certain systems?
- NIST MT Eval solution:
  - BLEU as the "official" metric, but report results with other metrics as well, and do some human judgment assessment
  - Best system according to BLEU was NOT always the best system according to human judgments
- WMT-09 Solution:
  - Human Preferences is the official metric
  - Systems can tune to whatever metric they want

Comparative MT Evaluations: Challenges and Pitfalls

- How meaningful are these comparative evaluations? What do we really learn?
  - They control for common training and test material, but much is left intentionally uncontrolled
    - Computational resources and runtime conditions
    - Various parameter settings within the systems
- How meaningful are these metric scores?
  - What does a BLEU (or METEOR) score of 30 or 50 mean?
  - Non-experts (i.e. DARPA higher ups) tend to think these scales are linear and extrapolate!
  - Multiple DARPA PMs have gotten into trouble because of this…

Why Not Use METEOR?

- METEOR is in many respects clearly superior to BLEU, so why hasn't it replaced BLEU so far?
  - Problems with using METEOR for MERT
    - Speed issues, length issues
    - SMT groups want to be tested using the metric to which they tune! But is this a good or bad thing?
  - No broad "buy-in" from the research community
    - "Everyone uses BLEU", BLEU is still the official metric for NIST MT Evals, so why use anything else?
    - Reuse of existing components such as MERT
    - Most MT researchers are not experts in MT evaluation, don't quite understand why using BLEU isn't good enough
  - Very limited amounts of research funding!
  - Publicity and promotion are difficult
  - Strong party interests, politics and rivalry

Some Consequences for MT Research

- Many MT research papers get accepted based on +1 point BLEU improvements
- Real error analysis of MT output is already very lacking…
- I worry NOT about the cases where BLEU went up… but rather about the cases where BLEU stayed the same or went slightly down
- I tell my students:
  - Run BLEU, METEOR and TER… if all go up, great!
  - If one goes up but the other goes down or stays the same, then something interesting is going on here!

Summary

- MT Evaluation is extremely important for driving system development and MT technology as a whole
- Human evaluations are costly, but are still the most meaningful
- There is no real substitute to human error analysis
- New "heavier" metrics that achieve better correlation with human judgments are being developed…
- But these are not yet practical for MERT
- Forums such as NIST Metrics MATR are new, but are very important to progress
- Methodology dilemmas in executing both micro and macro MT evaluations
- Lots of interesting and challenging work to do!

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?