Evaluating the Output of Machine Translation Systems

Alon Lavie

Associate Research Professor, Carnegie Mellon University

President, Safaba Translation Solutions

President, Association for MT in the Americas (AMTA)

AMTA 2010 Tutorial

Denver, Colorado, USA

October 31, 2010

Outline

- Motivation and Tutorial Goals
- Usage Scenarios: Important Distinctions
- MT Evaluation: Challenges, Dimensions and Approaches
- Human Evaluation Measures for MT
  - Case-Study: WMT-2009 Human Evaluation
- Automated Metrics for MT
  - BLEU, METEOR and TER
- Evaluating Automated Metrics for MT
  - Case-Study: NIST Metrics MATR 2008 Evaluation
- Usage Scenarios: In Practice
- Gaps and Summary

MT Evaluation

Why should you be interested?

- Practitioners and Users:
  - MT technology increasingly used within the industry
  - Increasing range of alternative systems, choices for building and customizing systems using outside vendors or in-house
  - How do you assess how well these alternatives perform, whether they are up to the tasks, whether they improve over time due to customization and further development?
  - Need for concrete measures for making informed decisions on investment, for calculating ROIs, and for quantifying the effectiveness of the alternatives you are considering
- Researchers:
  - MT Evaluation is a challenging and active research area of its own merit
  - Automated MT evaluation metrics are critical to state-of-the-art SMT

MT Evaluation

- Tutorial Goals:
  - Identify the most important usage scenarios for MT evaluation and the important distinctions between them
  - Provide you with a broad overview of the major state-of-the-art methods for human evaluation of MT output and automated metrics for MT evaluation
  - Expose you to the major issues involved in evaluating MT systems using both automated metrics and human assessment measures
  - Outline some of the major gaps and challenges, particularly within commercial settings

Translation Quality vs. MT Quality

- Quality assessment of translations commonly used within the industry (i.e. TEP process):
  - Every segment has to be translated correctly!
  - Quality measured by number of words edited/corrected in the editing (E) and/or proof-reading (P) stages
- Applying these same methods directly to the "raw" output of MT is usually not a meaningful endeavor:
  - MT requires some human post-editing to achieve human-level quality
  - The error profile exhibited by MT is very different than humans
  - Need for different types of evaluation measures:
    - Concrete measures for comparing/contrasting imperfect MT system performance
    - DO ASSESS whether MT improves productivity, and quantify improvement
    - DO ASSESS the quality of the resulting end human translation

Usage Scenarios: Important Distinctions

- Most Important Distinction:
  - **Offline "benchmark" testing of MT engine performance:**
    - Sample representative test documents with reference human translations are available
    - Commonly referred to as **Reference-based MT Evaluation**
  - **Operational Quality Assessment at runtime:**
    - MT engine is translating new source material
    - Need to identify whether the output is sufficient good for the underlying application (i.e. to pass along to human post-editors)
    - Commonly referred to as **Reference-less MT Confidence Scores**

Usage Scenarios: Important Distinctions

- Common Usage Scenarios for Reference-based Eval:
  - Compare performance of two or more different MT engines/technology for the same language-pair
  - Compare MT engine performance for two versions of the same engine/technology
    - Before and after customizing the engine
    - Before and after incremental development of the engine
  - Compare MT engine performance across different domains or types of input data
  - Compare MT engine performance on different sentence types, linguistic structures, other data distinctions

Usage Scenarios: Important Distinctions

- Common Usage Scenarios for MT Confidence Scores:
  - Identifying and flagging/filtering poorly translated segments during MT engine operation
  - Comparing alternative MT engines/technology in terms of their Quality Assessment capabilities and variation
    - Can the engines provide reliable Confidence Scores at runtime?
    - Segment Distributions: fraction of segments that pass Confidence Score thresholds
    - Example: what's better: Engine-1 with many "OK" translations and very few "Very Bad", or Engine-2 with many "Excellent" translations but equally many "Very Bad"?

Outline

- Motivation and Tutorial Goals
- Usage Scenarios: Important Distinctions
- MT Evaluation: Challenges, Dimensions and Approaches
- Human Evaluation Measures for MT
  - Case-Study: WMT-2009 Human Evaluation
- Automated Metrics for MT
  - BLEU, METEOR and TER
- Evaluating Automated Metrics for MT
  - Case-Study: NIST Metrics MATR 2008 Evaluation
- Usage Scenarios: In Practice
- Gaps and Summary

MT Evaluation: Major Issues

- MT Evaluation is Difficult:
  - Language variability – there is no single correct translation
  - Human evaluation is subjective
  - How good is “good enough”?
  - Is system A better than system B?
  - Depends on the target application and context
    - For what purpose will the MT output be used?
- Some well-established methods, but no standard or single approach that is universally accepted
- MT Evaluation is still a research topic in itself!
  - How do we assess whether an evaluation method is good?

Dimensions of MT Evaluation

- Human evaluation vs. automated metrics
- Quality assessment at sentence (segment) level vs. system level vs. task-based evaluation
- "Black-box" vs. "Glass-box" evaluation
- Evaluation for external validation vs. target function for automatic system tuning vs. ongoing quality assessment of MT output

Outline

- Motivation and Tutorial Goals
- Usage Scenarios: Important Distinctions
- MT Evaluation: Challenges, Dimensions and Approaches
- Human Evaluation Measures for MT
  - Case-Study: WMT-2009 Human Evaluation
- Automated Metrics for MT
  - BLEU, METEOR and TER
- Evaluating Automated Metrics for MT
  - Case-Study: NIST Metrics MATR 2008 Evaluation
- Usage Scenarios: In Practice
- Gaps and Summary

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 relatively crude and at times biased
  - Automatic metrics often 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

- Time and Cost
- Reliability and Consistency: difficulty in obtaining high-levels of intra and inter-coder agreement
- Developing meaningful statistical measures based on the collected human judgments
  - Example: if you collect information about the number, duration, and types of post editing operations, how do these translate into a global performance measure for the MT system?

Main Types of Human Assessments

- Adequacy and Fluency scores
- Human 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?

WMT-2009 MT Evaluations

- WMT-2009: 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-2009, including detailed analysis paper by organizers

Human Rankings at WMT-2009

- **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.

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

Human Editing at WMT-09

- Two Stages:
  - Humans edit the MT output to make it as fluent as possible
  - Judges evaluate the **edited output** for adequacy (meaning) with a **binary** Y/N judgment
- Instructions:
  - **Step-1:** Correct the translation displayed, making it as fluent as possible. If no corrections are needed, select “No corrections needed.” If you cannot understand the sentence well enough to correct it, select “Unable to correct.”
  - **Step-2:** Indicate whether the edited translations represent fully fluent and meaning equivalent alternatives to the reference sentence. The reference is shown with context, the actual sentence is bold.

Editing Interface

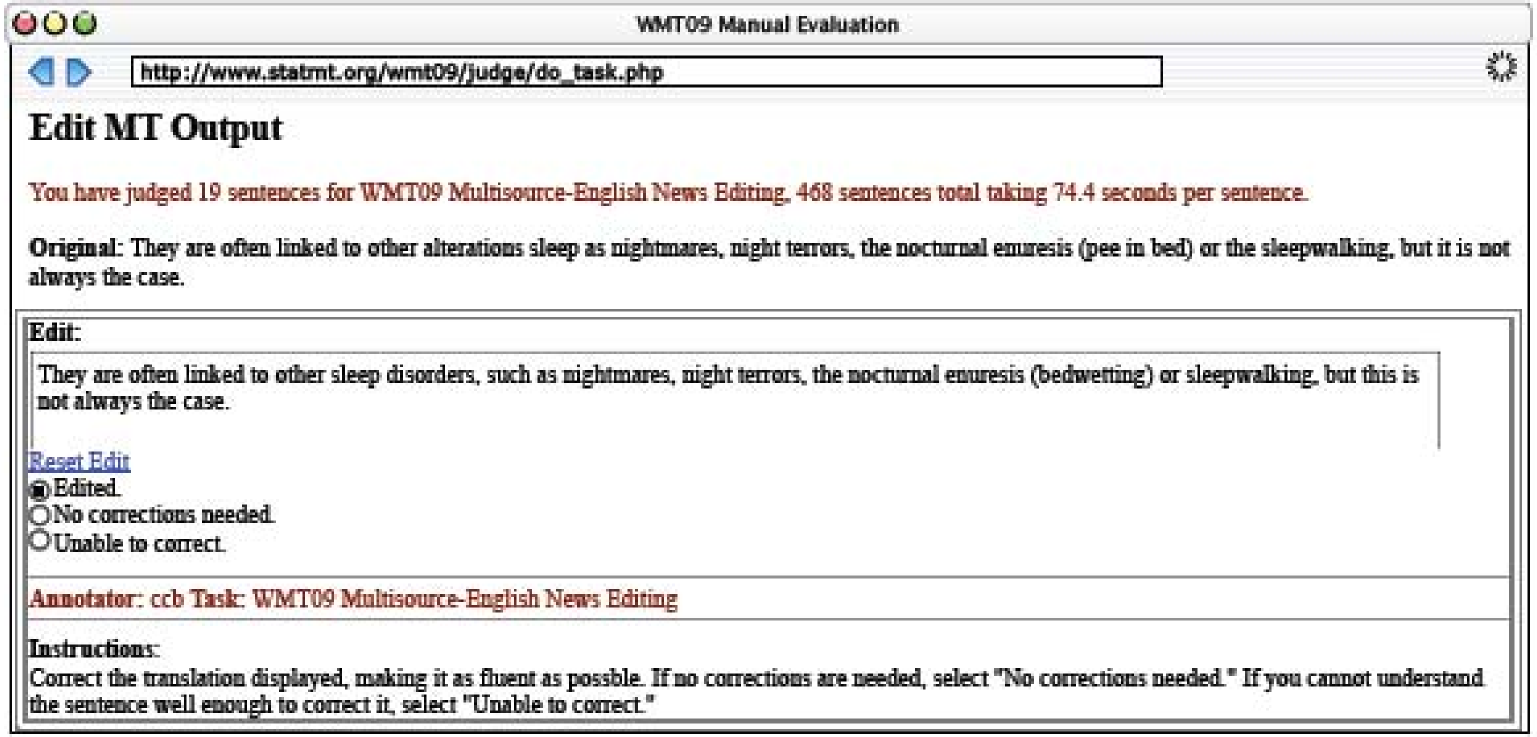
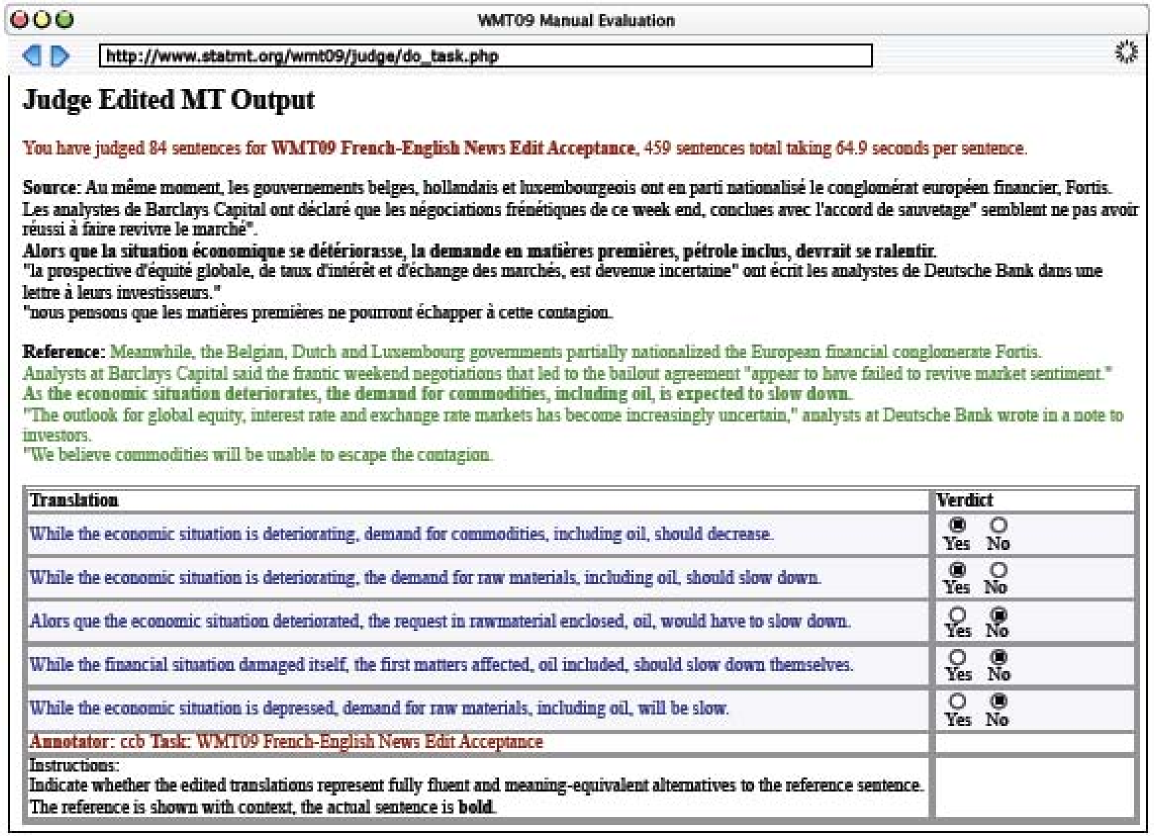


Figure 2: This screenshot shows an annotator editing the output of a machine translation system.

Evaluating Edited Output



WMT09 Manual Evaluation

http://www.statmt.org/wmt09/judge/do_task.php

Judge Edited MT Output

You have judged 84 sentences for **WMT09 French-English News Edit Acceptance**, 459 sentences total taking 64.9 seconds per sentence.

Source: Au même moment, les gouvernements belges, hollandais et luxembourgeois ont en parti nationalisé le conglomérat européen financier, Fortis.
Les analystes de Barclays Capital ont déclaré que les négociations frénétiques de ce week end, conclues avec l'accord de sauvetage" semblent ne pas avoir réussi à faire revivre le marché".
Alors que la situation économique se détériorasse, la demande en matières premières, pétrole inclus, devrait se ralentir.
"la prospective d'équité globale, de taux d'intérêt et d'échange des marchés, est devenue incertaine" ont écrit les analystes de Deutsche Bank dans une lettre à leurs investisseurs."
"nous pensons que les matières premières ne pourront échapper à cette contagion.

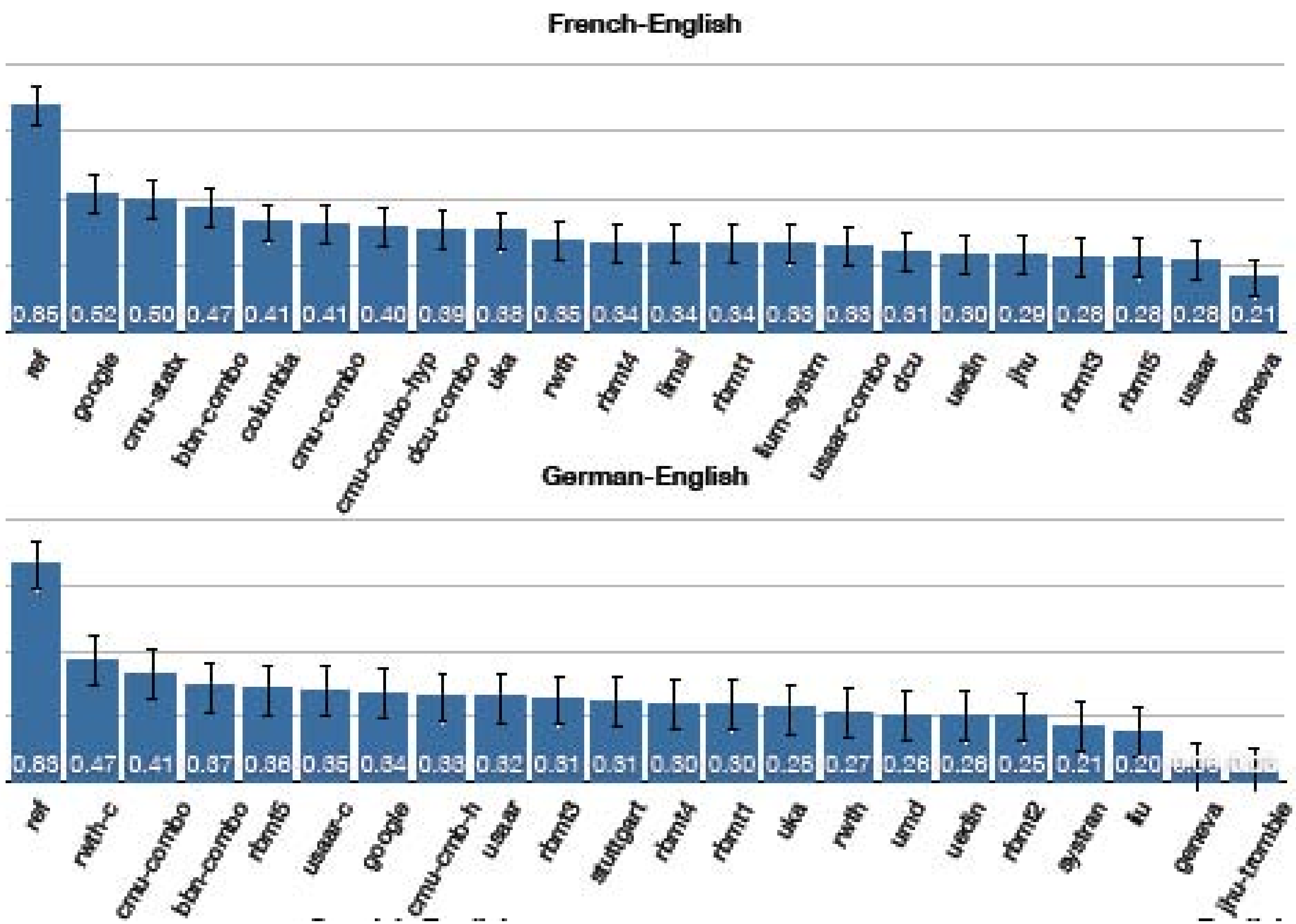
Reference: Meanwhile, the Belgian, Dutch and Luxembourg governments partially nationalized the European financial conglomerate Fortis.
Analysts at Barclays Capital said the frantic weekend negotiations that led to the bailout agreement "appear to have failed to revive market sentiment."
As the economic situation deteriorates, the demand for commodities, including oil, is expected to slow down.
"The outlook for global equity, interest rate and exchange rate markets has become increasingly uncertain," analysts at Deutsche Bank wrote in a note to investors.
"We believe commodities will be unable to escape the contagion.

Translation	Verdict
While the economic situation is deteriorating, demand for commodities, including oil, should decrease.	(●) Yes ( ) No
While the economic situation is deteriorating, the demand for raw materials, including oil, should slow down.	(●) Yes ( ) No
Alors que the economic situation deteriorated, the request in rawmaterial enclosed, oil, would have to slow down.	( ) Yes (●) No
While the financial situation damaged itself, the first matters affected, oil included, should slow down themselves.	( ) Yes (●) No
While the economic situation is depressed, demand for raw materials, including oil, will be slow.	( ) Yes (●) No
Annotator: ccb **Task:** WMT09 French-English News Edit Acceptance	
Instructions: Indicate whether the edited translations represent fully fluent and meaning-equivalent alternatives to the reference sentence. The reference is shown with context, the actual sentence is **bold**.	

Figure 3: This screenshot shows an annotator judging the acceptability of edited translations.

Human Editing Results

- **Goal:** to assess how often a systems MT output is "fixable" by a human post-editor
- **Measure used:** fraction of time that humans assessed that the edited output had the same meaning as the reference


French-English
0.85 0.52 0.50 0.47 0.41 0.41 0.40 0.39 0.38 0.35 0.34 0.34 0.34 0.33 0.33 0.31 0.30 0.29 0.28 0.28 0.28 0.21
ref
google
cmu-statx
bbn-combo
columbia
cmu-combo
cmu-combo-hyp
dcu-combo
uka
rwth
rbmt4
limsi
rbmt1
tum-systrn
usaar-combo
dcu
uedin
jhu
rbmt3
rbmt5
usaar
geneva
German-English
0.83 0.47 0.41 0.37 0.36 0.35 0.34 0.33 0.32 0.31 0.31 0.30 0.30 0.28 0.27 0.26 0.26 0.25 0.21 0.20
ref
rwth-c
cmu-combo
bbn-combo
rbmt5
usaar-c
google
cmu-cmb-h
usaar
rbmt3
stuttgart
rbmt4
rbmt1
uka
rwth
umd
uedin
rbmt2
systran
liu
geneva
jhu-tromble

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

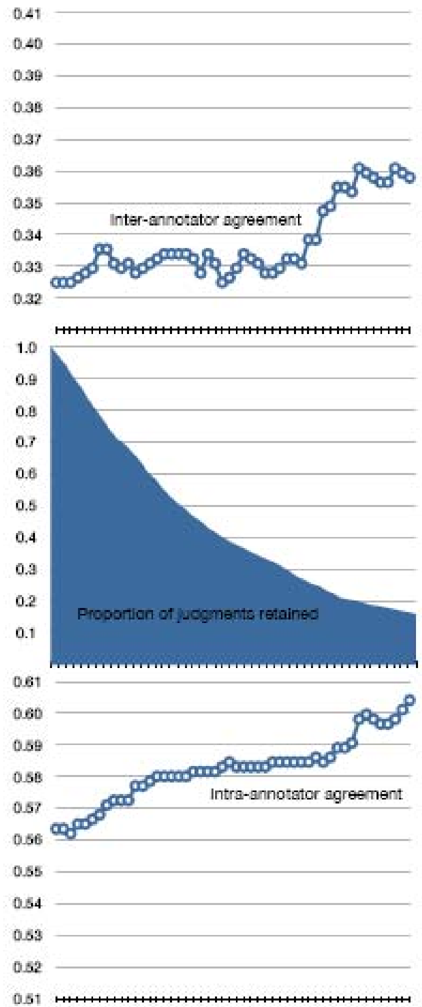


Figure 4: The effect of discarding every annotators' initial judgments, up to the first 50 items

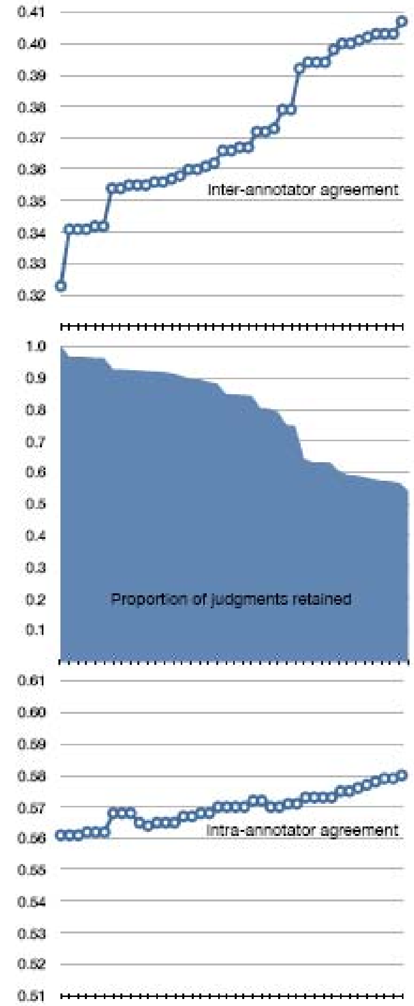


Figure 5: The effect of removing annotators with the lowest agreement, disregarding up to 40 annotators

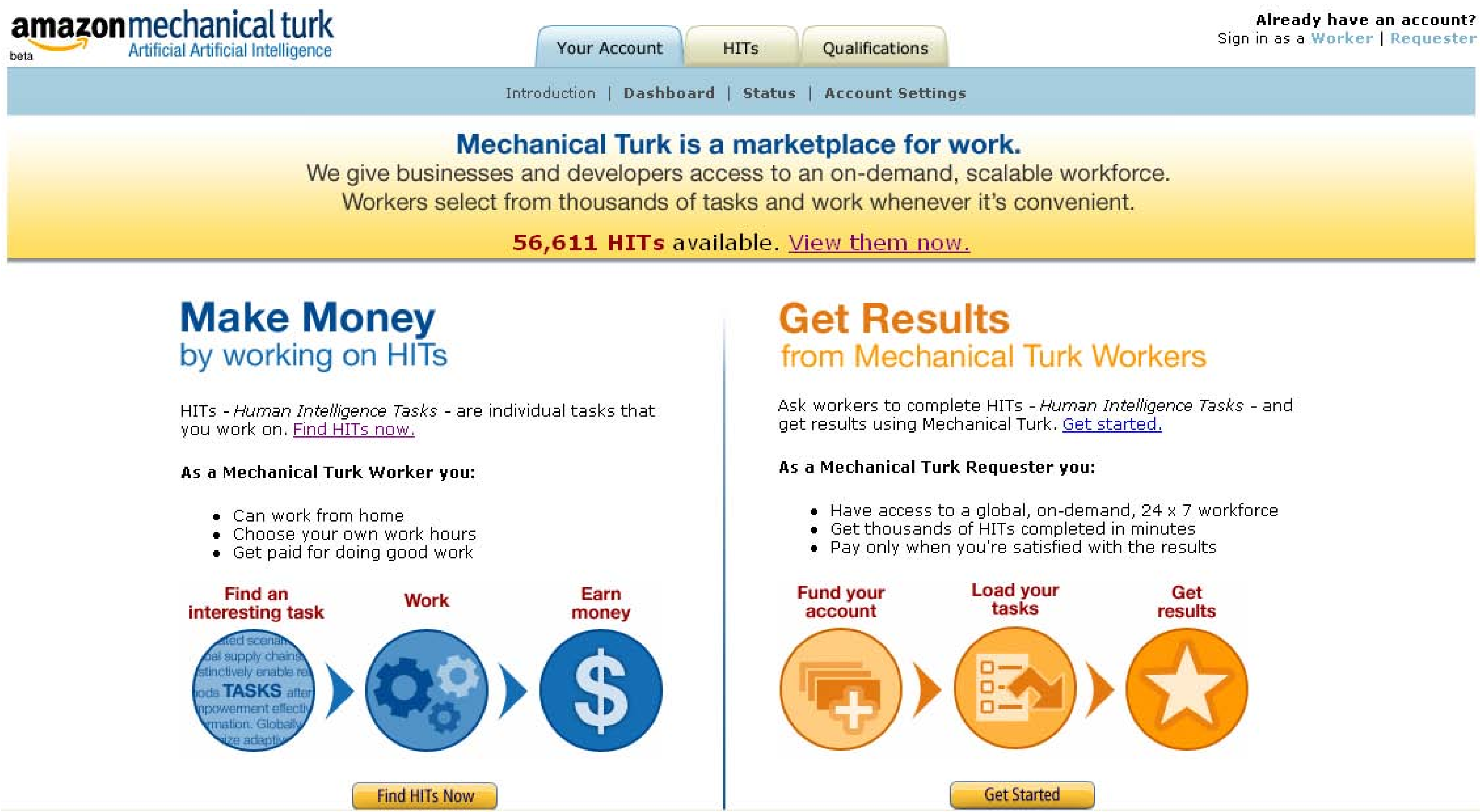
Cost and Quality Issues

- **High cost** and **controlling for agreement quality** are the most challenging issues in conducting human evaluations of MT output
- Critical decisions:
  - Your human judges: professional translators? Non-expert bilingual speakers? Target-language only speakers?
  - Where do you recruit them? How do you train them?
  - How many different judgments per segment to collect?
  - Easy to overlook issues (i.e. the user interface) can have significant impact on quality and agreement
- Measure intra- and inter-coder agreement as an integral part of your evaluation!

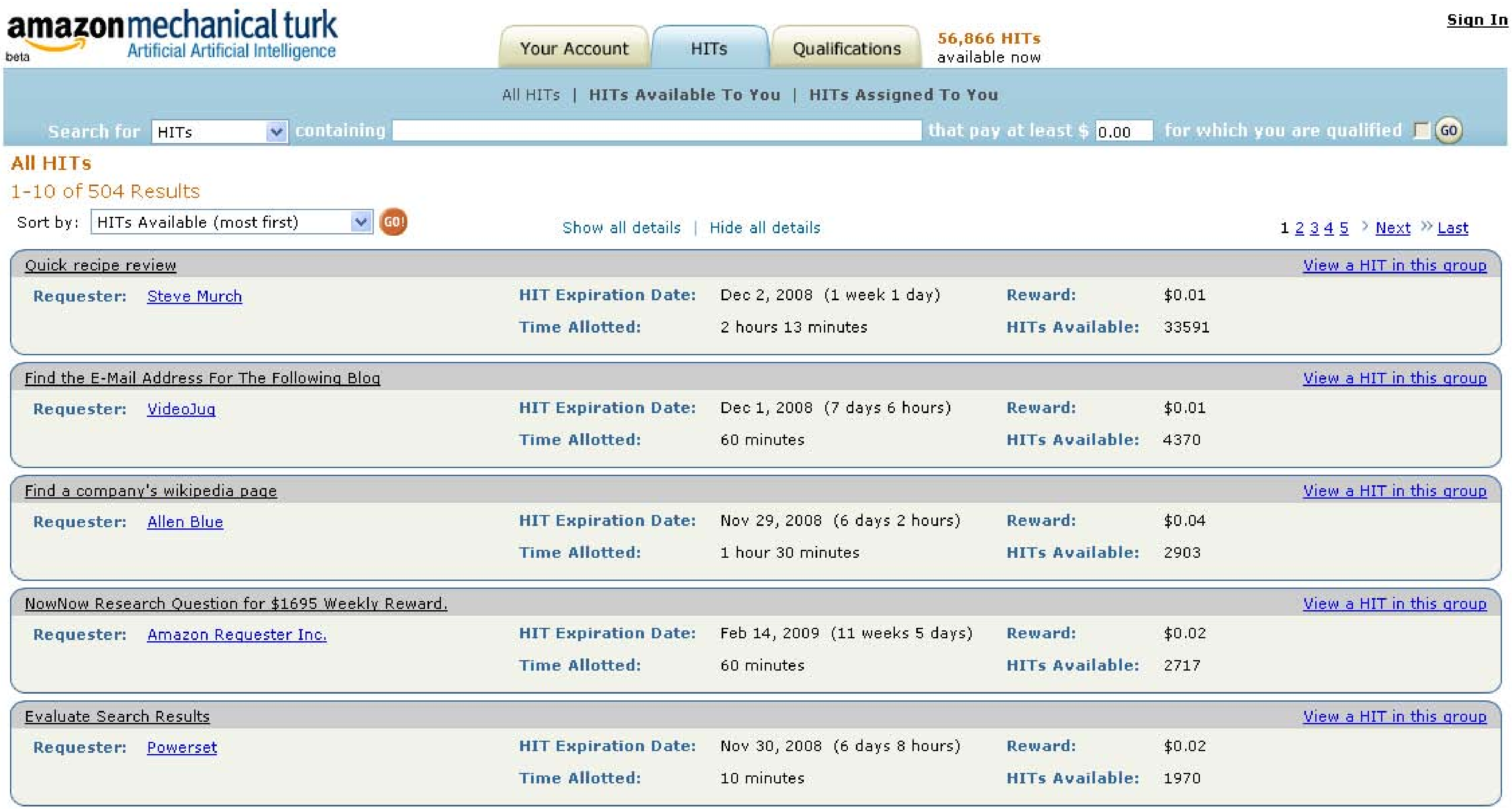
Human Evaluations Using Crowd-Sourcing

- Recent popularity of crowd-sourcing has introduced some exciting new ideas for human assessment of MT output
  - Using the “crowd” to provide human judgments of MT quality, either directly or indirectly
  - Amazon’s Mechanical Turk as a labor source for human evaluation of MT output

Mechanical Turk



Mechanical Turk



HIT Preview

Rate this translation (قم بتقيم نوعية الترجمة)

Instructions (English): Below are two translations of the same English sentence into Arabic. The first was written by a human translator and the second was translated automatically by a computer. Please rate the extent to which the automatic translation has the same meaning as the human translation.

تعليمات : أدناه مُعطى ترجمتان بالعربية لنفس الجملة الإنكليزية . الترجمة الاولى تمت على يد مترجم بشري بينما الثانية تمت أوتوماتيكيا بواسطة كمبيوتر. رجاءً قم بتقيم مدى توافق معنى الترجمة الأوتوماتيكية مع معنى الترجمة البشرية

Scale and Examples:

Score (التقيم):	Human Translation (ترجمة بشرية) Automatic Translation (ترجمة آلية)
4 - Excellent (ممتاز):	واكد موسيفيني على حاجة الكوميسا والدول الافريقية الى الاتحاد ، حتى تحصل على فرصة افضل فى عالم العولمة موسيفيني شدد على الحاجة إلى دول الكوميسا والدول الافريقية الى التوحد من أجل منحهم فرصة أفضل فى عالم العولمة.
3 - Good (جيد):	وستبلغ القيمة المضافة للصناعة 328 مليار يوان بزيادة 12 بالمئة وقيمة الصادرات مئة مليار دولار امريكى بزيادة 8 بالمئة. في هذه الصناعة ذات القيمة المضافة ستكون 328 مليار يوان ، بزيادة 12 ٪ ، بينما لرتفعت الصادرات ستصل إلى 100 مليون دولار ، أي بزيادة 8٪.
2 - Bad (سيئة):	الا انه لم يتم فعلا تقديم سوى 7.17 مليون فقط. ولكن فقط 17.7 مليون الواردة في الواقع.
1 - Very bad (سيئة جدا):	جائزة النقاد العرب في مهرجان كان لفيلم ;يا ولاد; للمخرج زياد الدويري النقاد العرب 'على جائزة في مهرجان كان السينمائي يذهب إلى; بيروت الغربية لزياد دويري

Task:

Human translation (ترجمة بشرية):	ستة فنان من 61 دولة يشاركون في أول معرض رسمي مصري للرسم على البورسلين
Automatic translation (ترجمة آلية):	فنانا من 16 دولة تشارك في أول البورسلين المصرية معرض التصوير 100
Rating (التقيم):	◯ 4 - Excellent (ممتازة) ◯ 3 - Good (جيدة) ◯ 2 - Bad (سيئة) ◯ 1 - Very bad (سيئة جدا)

Please provide any comments you may have below, we appreciate your input!
رجاءً قم بتقيم أية ملاحظات قد تكون لديك فنحن

Submit

Outline

- Motivation and Tutorial Goals
- Usage Scenarios: Important Distinctions
- MT Evaluation: Challenges, Dimensions and Approaches
- Human Evaluation Measures for MT
  - Case-Study: WMT-2009 Human Evaluation
- **Automated Metrics for MT**
  - **BLEU, METEOR and TER**
- Evaluating Automated Metrics for MT
  - Case-Study: NIST Metrics MATR 2008 Evaluation
- Usage Scenarios: In Practice
- Gaps and Summary

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 still relatively crude, do not distinguish well between subtle differences in systems
  - Individual sentence scores are often not very reliable, aggregate scores on a large test set are more stable
- Automated metrics for MT evaluation are still a very active area of current research

Desirable Automated 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

Automated Metric Components

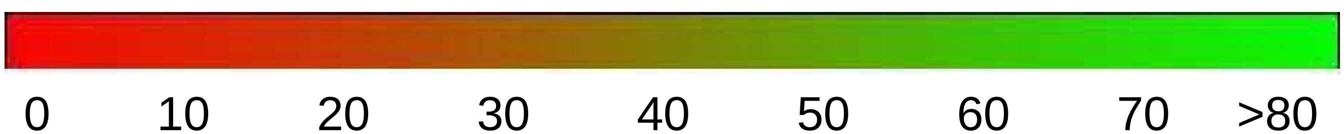
- Example:
  - Reference: "the Iraqi weapons are to be handed over to the army within two weeks"
  - MT output: "in two weeks Iraq's weapons will give army"
- Possible metric components:
  - Precision: correct words / total words in MT output
  - Recall: correct words / total words in reference
  - Combination 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,

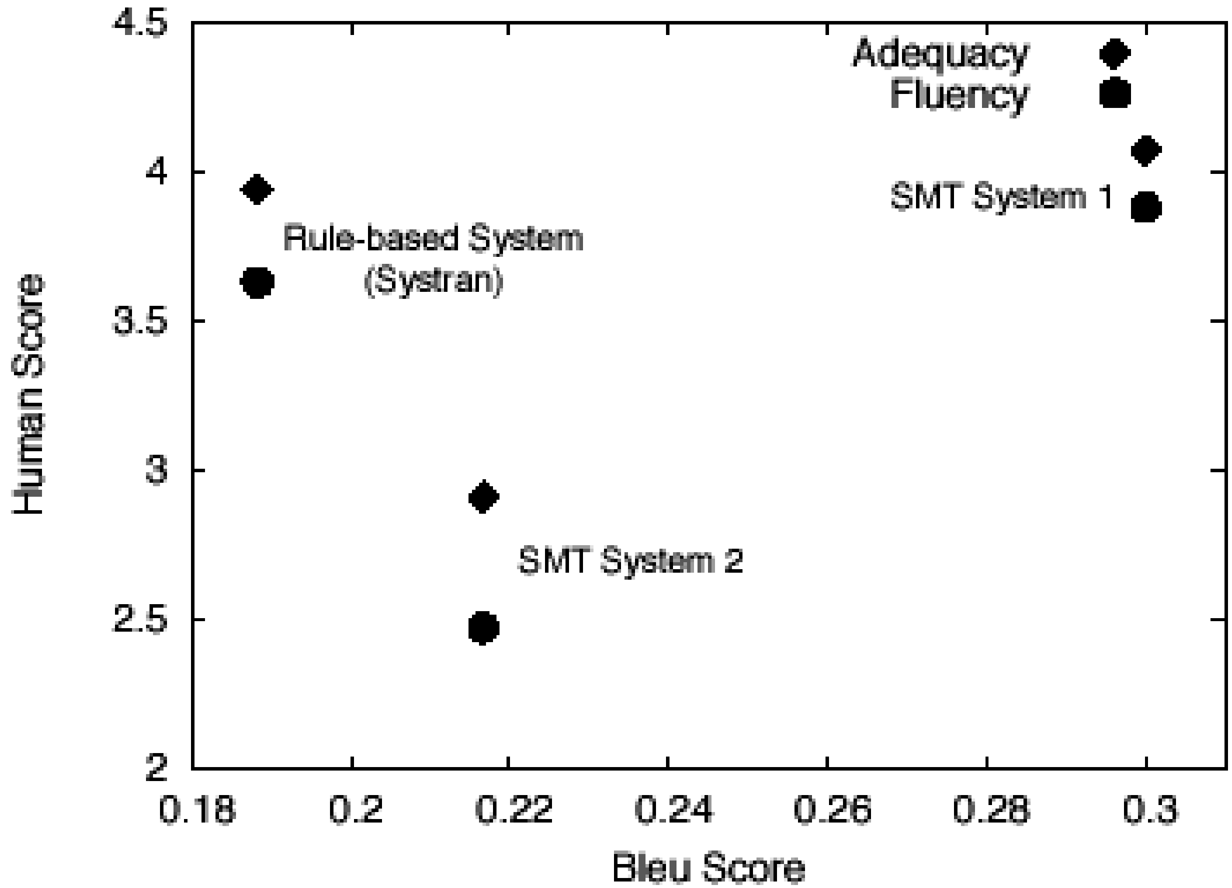
$$\textsc{Bleu} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) .$$

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

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



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, synonyms and paraphrases)
  - 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 different types of human judgments for each language
- 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, single and multi word paraphrases
  - 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
- All previous versions of METEOR used a greedy three-stage matching: exact, stem, synonyms
- New version of METEOR uses an integrated one stage search

Matcher Example

the sri lanka prime minister criticizes the leader of the country

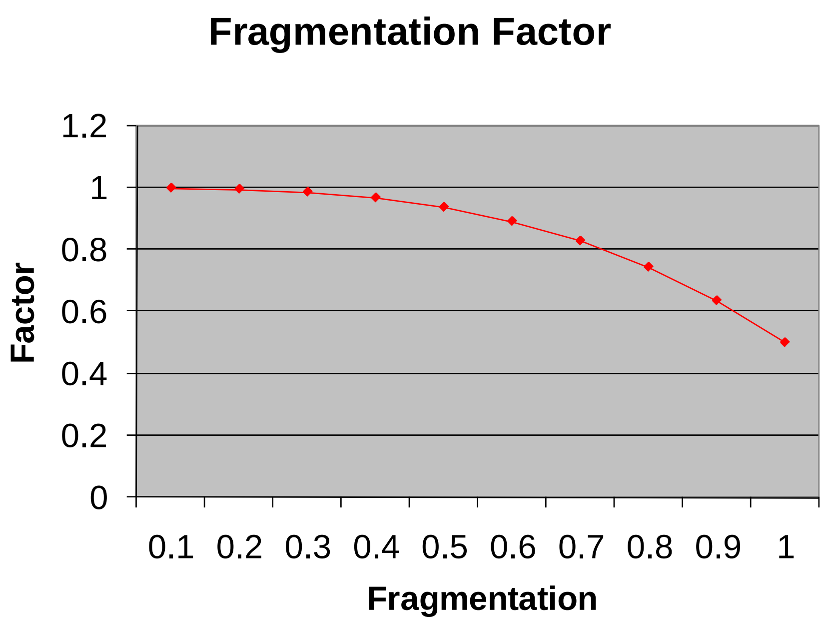
President of Sri Lanka criticized by the country’s Prime Minister

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)

The 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:

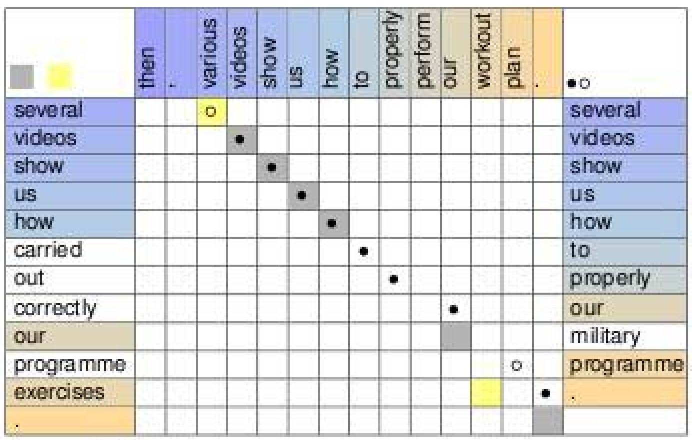
Fmean * (1- DF) = 0.3731 * 0.9375 = 0.3498

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

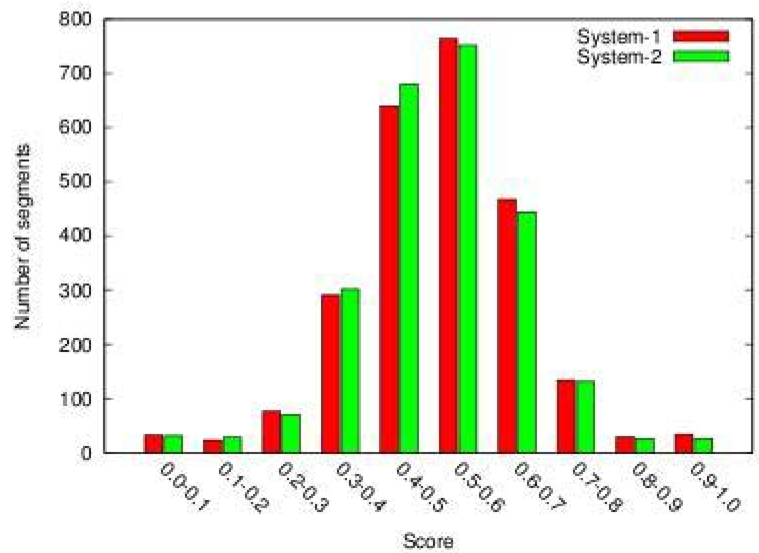
METEOR Analysis Tools

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



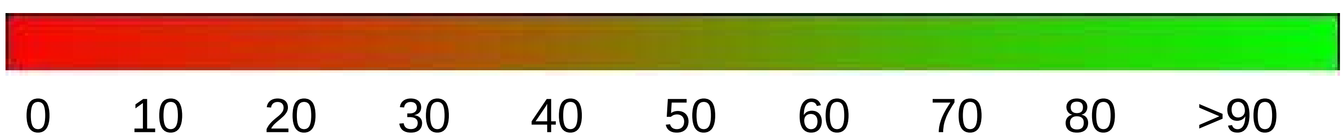
Segment 2001

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

Practical Notes of Use for Automated Metrics

- BLEU and METEOR are freely available for commercial use, TERp is NOT (unsure about TER)
- Symantec has an evaluation suite tool (SymEval) that allows comparing MT output before and after human post-editing with GTM and other scores – will be releasing it Open Source soon [based on personal communication with Johann Roturier]
- Asia Online's Language Studio Lite has a freely available evaluation suite tool that supports easy evaluation using BLEU, F-Measure and TER

MT Confidence Scores

- Difficult problem, but of significant importance to MT usage within the commercial translation industry
- Recent work on this problem has shown some encouraging success
  - Work by [Specia et. al. 2010] on developing a multi-feature classifier for producing MT confidence scores
  - Language Weaver now produces a confidence measure that is returned with each translation
- These scores can be used to filter out poor MT-produced translations, so that they are not sent to post-editing

Outline

- Motivation and Tutorial Goals
- Usage Scenarios: Important Distinctions
- MT Evaluation: Challenges Dimensions and Approaches
- Human Evaluation Measures for MT
  - Case-Study: WMT-2009 Human Evaluation
- Automated Metrics for MT
  - BLEU, METEOR and TER
- **Evaluating Automated Metrics for MT**
  - Case-Study: NIST Metrics MATR 2008 Evaluation
- Usage Scenarios: In Practice
- Gaps and Summary

Comparing Metrics

- 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

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

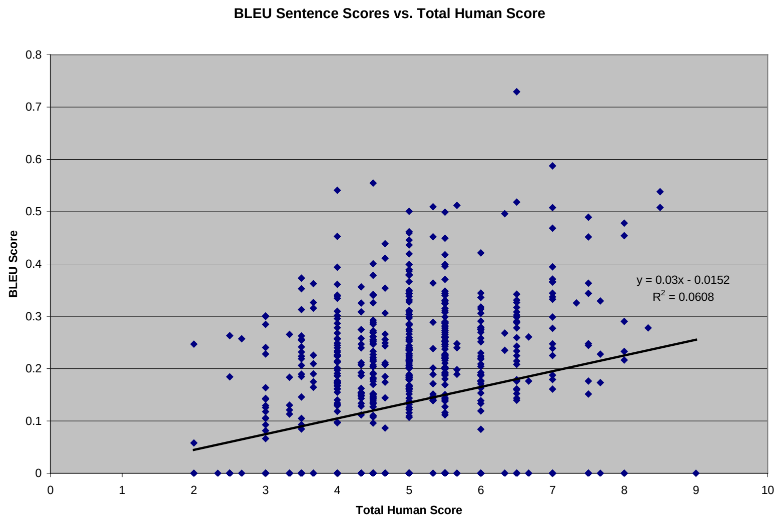
- Human Assessment Type: **Preferences, Pair-wise comparison across systems**
- Target Language: **English**
- Correlation Level: **segment**

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)

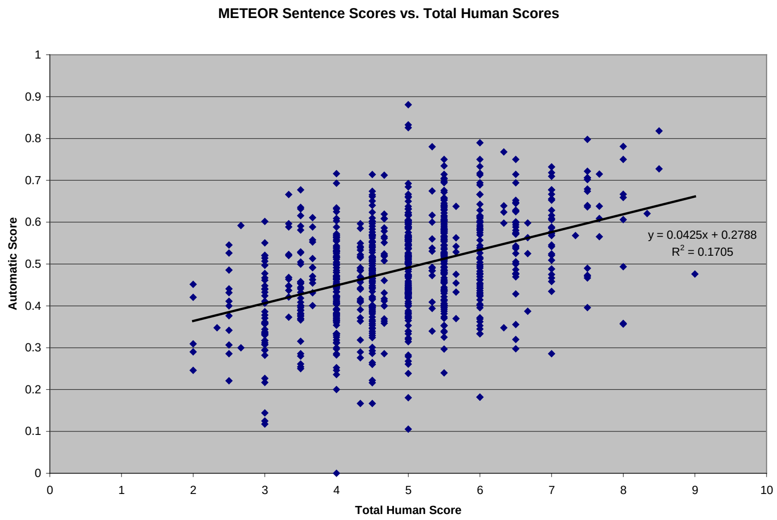
METEOR vs. BLEU
Sentence-level Scores
(CMU SMT System, TIDES 2003 Data)

R=0.2466



BLEU

R=0.4129



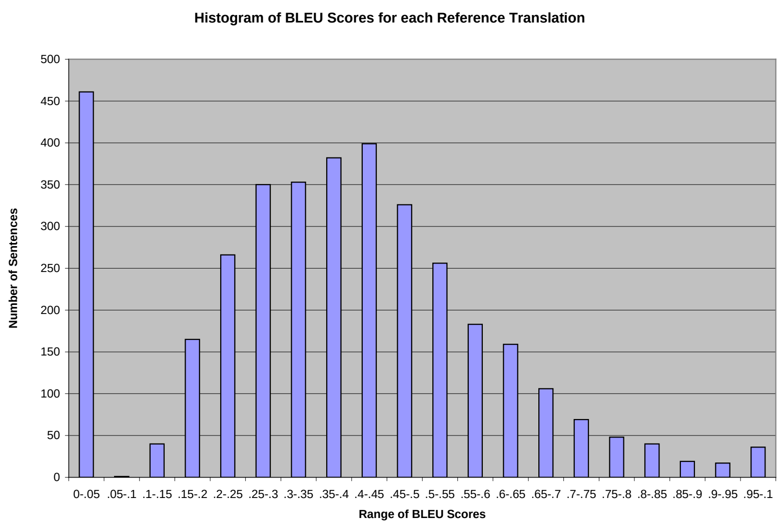
METEOR

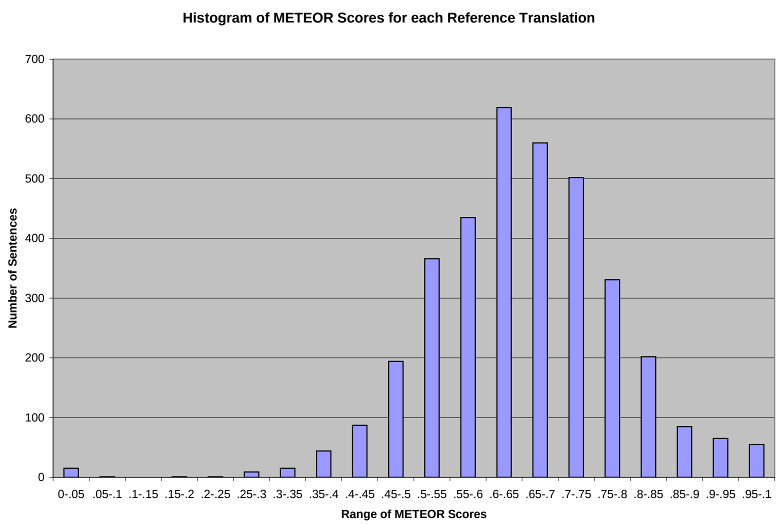
METEOR vs. BLEU

Histogram of Scores of Reference Translations
2003 Data

Mean=0.3727 STD=0.2138

Mean=0.6504 STD=0.1310





BLEU

METEOR

Outline

- Motivation and Tutorial Goals
- Usage Scenarios: Important Distinctions
- MT Evaluation: Challenges, Dimensions and Approaches
- Human Evaluation Measures for MT
  - Case-Study: WMT-2009 Human Evaluation
- Automated Metrics for MT
  - BLEU, METEOR and TER
- Evaluating Automated Metrics for MT
  - Case-Study: NIST Metrics MATR 2008 Evaluation
- **Usage Scenarios: In Practice**
- Gaps and Summary

Comparing MT Systems

- Scenario:
  - Compare several alternative available MT engines for a specific client or domain
  - Compare a system before and after significant MT system customization for a specific client or domain
- Approach:
  - Select and prepare a meaningful evaluation set along with a human reference translation (at least one)
    - Set of documents representative of client data that was NOT USED for MT system development or tuning
    - Evaluation data can often be extracted from client's existing TMs, but make sure these are clean and formatted for running MT metrics
  - Run all three major metrics: BLEU, METEOR and TER

Tuning an SMT System

- Scenario:
  - Need to tune the parameters of a newly trained SMT system (such as Moses) for a specific client or domain
- Approach:
  - Create a tuning data set, representative of the client data or domain, which was NOT USED for system development, along with a human reference translation (preferably more than one)
  - BLEU is the most commonly used metric for tuning (some implementations REQUIRE using BLEU)
  - Tuning with BLEU is most stable if the set is at least 500 segments and has four reference translations

Task-based Assessment

- Scenario:
  - Assessing whether post-editing MT output is cost effective for a specific MT system and client or domain
- Approach:
  - Be aware that the specific setup of how MT is integrated within the translation process is critical
  - Create a segment-level quality profile using METEOR or TER
  - You will likely want/need to conduct a human study where you actually measure translation cost and time with MT post-editing, and compare with a baseline of not using MT at all
  - Leverage your client TMs as much as possible
  - If possible, use confidence scores to filter out poor MT segments

Outline

- Motivation and Tutorial Goals
- Usage Scenarios: Distinctions
- MT Evaluation: Challenges, Dimensions and Approaches
- Human Evaluation Measures for MT
  - Case-Study: WMT-2009 Human Evaluation
- Automated Metrics for MT
  - BLEU, METEOR and TER
- Evaluating Automated Metrics for MT
  - Case-Study: NIST Metrics MATR 2008 Evaluation
- Usage Scenarios: In Practice
- Gaps and Summary

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)
- Need for more work on reference-less confidence scores for filtering poor MT (for post-editors and human translators)

Summary

- MT evaluation measures are critical for assessing the performance and ROI of MT systems in commercial settings
- Both human measures and automatic metrics are important, for different purposes
- If you are going to conduct a human evaluation, consult with an experienced expert or vendor
- If you are going to use automatic metrics, learn what they mean, how to interpret their scores, and which metric or measure is most suitable for your task

Acknowledgements

- NIST Metrics MATR data and results are courtesy of Mark Przybocki and the entire NIST MT evaluation group
- WMT-2009 data and results are courtesy of Chris Callison-Burch and the WMT workshop series organizing team
- Thanks to my students who have worked on various aspects of METEOR: Satanjeev Banerjee, Kenji Sagae, Abhaya Agarwal and Michael Denkowski

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?