MT Evaluation: Human Measures and Assessment Methods

11-731:
Machine Translation

Alon Lavie

February 12, 2013

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
- MT Evaluation is difficult:
  - Language Variability: there is no single correct translation
  - Human evaluation is subjective
  - How good is "good enough"? Depends on target 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 is good?

Dimensions of MT Evaluation

- Human evaluation vs. automatic 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. contrastive comparison of different MT systems vs. target function for automatic MT system tuning

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 rather crude and at times biased
  - Automatic metrics usually 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
  - **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 collected human judgments
  - Example: if collecting binary judgments for sentences, how do these map into global scores?

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 Post-Editing 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]
- Initiated during DARPA MT evaluations during mid-1990s
- Most commonly used until recently
- Main Issues: definitions of scales, agreement, normalization across judges

Human 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-2012

- WMT-2012: Shared task on developing MT systems between several European languages (to English and from English)
- Also included tracks on automated MT metric evaluation and quality estimation
- Official Metric: Human Rankings
- Detailed evaluation and analysis of results
- 2-day Workshop at NAACL-2012, including detailed analysis paper by organizers

Human Rankings at WMT-2012

- **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 all language pairs there were more than 5 system 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 are ranked based on the fraction of comparison instances for which they were judged to be better than any other system.

Assessing MT Systems

- Human Rankings were used to assess:
  - Which systems produced the best translation quality for each language pair?
  - Which of the systems that used only the provided training materials (“constrained”) produced the best translation quality?

Czech-English
3,603–3,718 comparisons/system

System	C?	>others
ONLINE-B •	N	0.65
UEDIN ⋆	Y	0.60
CU-BOJAR	Y	0.53
ONLINE-A	N	0.53
UK	Y	0.37
JHU	Y	0.32

Spanish-English
1,527–1,775 comparisons/system

System	C?	>others
ONLINE-A •	N	0.62
ONLINE-B •	N	0.61
QCRI ⋆	Y	0.60
UEDIN •⋆	Y	0.58
UPC	Y	0.57
GTH-UPM	Y	0.52
RBMT-3	N	0.51
JHU	Y	0.48
RBMT-4	N	0.46
RBMT-1	N	0.42
ONLINE-C	N	0.42
UK	Y	0.19

French-English
1,437–1,701 comparisons/system

System	C?	>others
LIMSI •⋆	Y	0.63
KIT •⋆	Y	0.61
ONLINE-A •	N	0.59
CMU •⋆	Y	0.57
ONLINE-B •	N	0.57
UEDIN	Y	0.55
LIUM	Y	0.52
RWTH	Y	0.52
RBMT-1	N	0.46
RBMT-3	N	0.46
UK	Y	0.44
SFU	Y	0.44
RBMT-4	N	0.43
JHU	Y	0.41
ONLINE-C	N	0.32

English-Czech
2,652–3,146 comparisons/system

System	C?	>others
CU-DEPFIX •	N	0.66
ONLINE-B	N	0.63
UEDIN ⋆	Y	0.56
CU-TAMCH	N	0.56
CU-BOJAR ⋆	Y	0.54
CU-TECTOMT ⋆	Y	0.53
ONLINE-A	N	0.53
COMMERCIAL-1	N	0.48
COMMERCIAL-2	N	0.46
CU-POOR-COMB	Y	0.44
UK	Y	0.44
SFU	Y	0.36
JHU	Y	0.32

English-Spanish
2,013–2,294 comparisons/system

System	C?	>others
ONLINE-B •	N	0.65
RBMT-3	N	0.58
ONLINE-A •	N	0.56
PROMT	N	0.55
UPC ⋆	Y	0.52
UEDIN ⋆	Y	0.52
RBMT-4	N	0.46
RBMT-1	N	0.45
ONLINE-C	N	0.43
UK	Y	0.41
JHU	Y	0.36

English-French
1,410–1,697 comparisons/system

System	C?	>others
LIMSI •⋆	Y	0.66
RWTH	Y	0.62
ONLINE-B	N	0.60
KIT •⋆	Y	0.59
LIUM	Y	0.55
UEDIN	Y	0.53
RBMT-3	N	0.52
ONLINE-A	N	0.51
PROMT	N	0.51
RBMT-1	N	0.48
JHU	Y	0.44
UK	Y	0.40
RBMT-4	N	0.39
ONLINE-C	N	0.39
ITS-LATL	N	0.36

German-English
1,386–1,567 comparisons/system

System	C?	>others
ONLINE-A •	N	0.65
ONLINE-B •	N	0.65
QUAERO	Y	0.61
RBMT-3	N	0.60
UEDIN ⋆	Y	0.60
RWTH ⋆	Y	0.56
KIT ⋆	Y	0.55
LIMSI	Y	0.54
QCRI	Y	0.52
RBMT-1	N	0.51
RBMT-4	N	0.50
ONLINE-C	N	0.43
DFKI-BERLIN	Y	0.40
UK	Y	0.37
JHU	Y	0.34
UG	Y	0.17

English-German
1,777–2,160 comparisons/system

System	C?	>others
ONLINE-B •	N	0.64
RBMT-3	N	0.63
RBMT-4 •	N	0.58
RBMT-1	N	0.56
LIMSI ⋆	Y	0.55
ONLINE-A	N	0.54
UEDIN-WILLIAMS ⋆	Y	0.51
KIT ⋆	Y	0.50
DFKI-HUNSICKER	N	0.48
UEDIN ⋆	Y	0.47
RWTH ⋆	Y	0.47
ONLINE-C	N	0.47
UK	Y	0.45
JHU	Y	0.43
DFKI-BERLIN	Y	0.25

C? indicates whether system is constrained (unhighlighted rows): trained only using supplied training data, standard monolingual linguistic tools, and, optionally, LDC's English Gigaword.

- • indicates a **win**: no other system is statistically significantly better at p-level ≤ 0.10 in pairwise comparison.
- ⋆ indicates a ***constrained* win**: no other *constrained* system is statistically better.

Table 4: Official results for the WMT12 translation task. Systems are ordered by their > others score, reflecting how often their translations won in pairwise comparisons. For detailed head-to-head comparisons, see Appendix A.

French-English

1,437–1,701 comparisons/system

System	C?	>others
LIMSI ●★	Y	0.63
KIT ●★	Y	0.61
ONLINE-A ●	N	0.59
CMU ●★	Y	0.57
ONLINE-B ●	N	0.57
UEDIN	Y	0.55
LIUM	Y	0.52
RWTH	Y	0.52
RBMT-1	N	0.46
RBMT-3	N	0.46
UK	Y	0.44
SFU	Y	0.44
RBMT-4	N	0.43
JHU	Y	0.41
ONLINE-C	N	0.32

Methods for Overall Ranking

- Different possible ways to calculate overall system rankings based on the collected segment-level ranking judgments
- WMT-2012 surveys six different possible methods and compares five of them on the data collected for English-German MT systems
- Different methods generate mostly but not fully similar results
- Statistical significance can be established based on the variance within the collected data, using bootstrap sampling

Methods for Overall Ranking

	Bojar	Lopez	Most Probable	MC Playoffs	Expected Wins
1	0.641: ONLINE-B	RBMT-4	RBMT-4	6.16: ONLINE-B	0.640 (1-2): ONLINE-B
2	0.627: RBMT-3	ONLINE-B	ONLINE-B	6.39: RBMT-3	0.622 (1-2): RBMT-3
3	0.577: RBMT-4	RBMT-3	RBMT-3	6.98: RBMT-4	0.578 (3-5): RBMT-4
4	0.557: RBMT-1	RBMT-1	RBMT-1	7.32: RBMT-1	0.553 (3-6): RBMT-1
5	0.547: LIMSI	ONLINE-A	ONLINE-A	7.46: LIMSI	0.543 (3-7): LIMSI
6	0.537: ONLINE-A	UEDIN-WILLIAMS	LIMSI	7.57: ONLINE-A	0.534 (4-8): ONLINE-A
7	0.509: UEDIN-WILLIAMS	LIMSI	UEDIN-WILLIAMS	7.87: UEDIN-WILLIAMS	0.511 (5-9): UEDIN-WILLIAMS
8	0.503: KIT	KIT	KIT	7.98: KIT	0.503 (6-11): KIT
9	0.476: DFKI-HUNSICKER	DFKI-HUNSICKER	DFKI-HUNSICKER	8.32: UEDIN	0.477 (7-13): UEDIN
10	0.475: UEDIN	ONLINE-C	ONLINE-C	8.38: DFKI-HUNSICKER	0.472 (8-13): DFKI-HUNSICKER
11	0.470: RWTH	UEDIN	UEDIN	8.41: ONLINE-C	0.470 (8-13): ONLINE-C
12	0.470: ONLINE-C	UK	UK	8.44: RWTH	0.468 (8-13): RWTH
13	0.448: UK	RWTH	RWTH	8.72: UK	0.447 (10-14): UK
14	0.435: JHU	JHU	JHU	8.87: JHU	0.434 (12-14): JHU
15	0.249: DFKI-BERLIN	DFKI-BERLIN	DFKI-BERLIN	11.15: DFKI-BERLIN	0.249 (15): DFKI-BERLIN

Table 5: Overall ranking with different methods (English–German)

Human Post-Editing

- A natural task-based evaluation measure for utility of MT output
  - Human translator(s) edit the output of the MT system into a correct translation
  - Measure the amount of “effort” involved
- Practical: increasing number of commercial translation agencies are actually doing MTPE
- Challenges:
  - How do you measure post-editing “effort”?
  - Large variations across translators – training is important
  - Bilingual translators are costly – can monolingual target-language speakers do this reliably?

TER

- Translation Edit (Error) Rate (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, but is not a true measure of post-editing effort

HTER

- Human Translation Edit Rate
- Developed as the official evaluation measure of the DARPA GALE program and continues to be used in BOLT
- Evaluation Process:
  - Team of translators post-edits the MT segment
  - TER is used to find the minimum-distance post-edited human reference
  - Aggregate system-level HTER scores are calculated at the document-level
  - Ranked document lists are generated for each system
  - Systems are scored based on fraction of documents that pass threshold levels of TER performance

Human Editing at WMT-2009

- 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

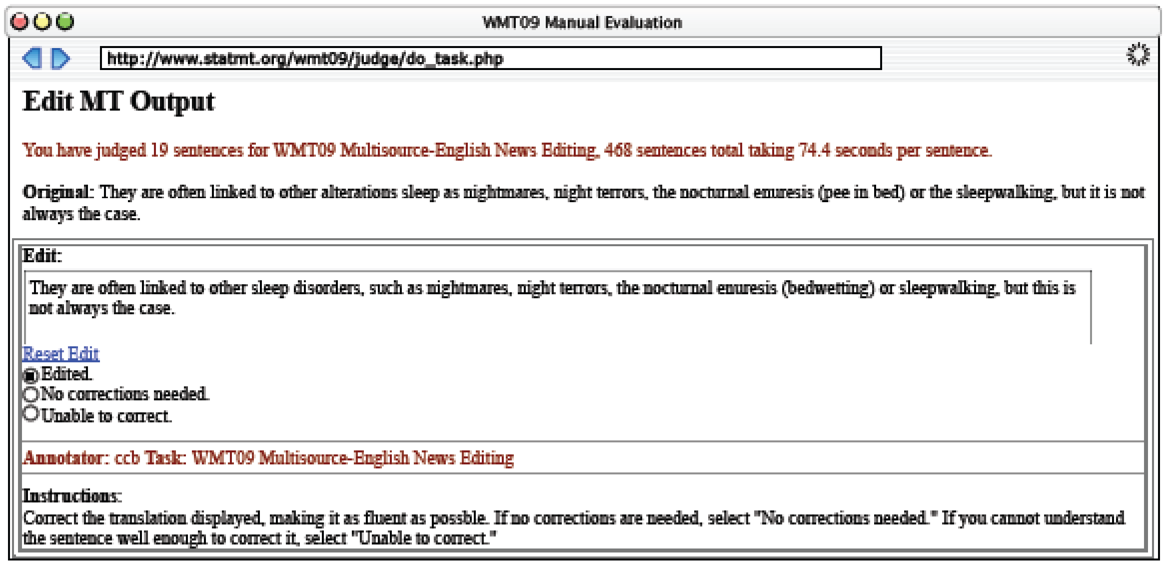
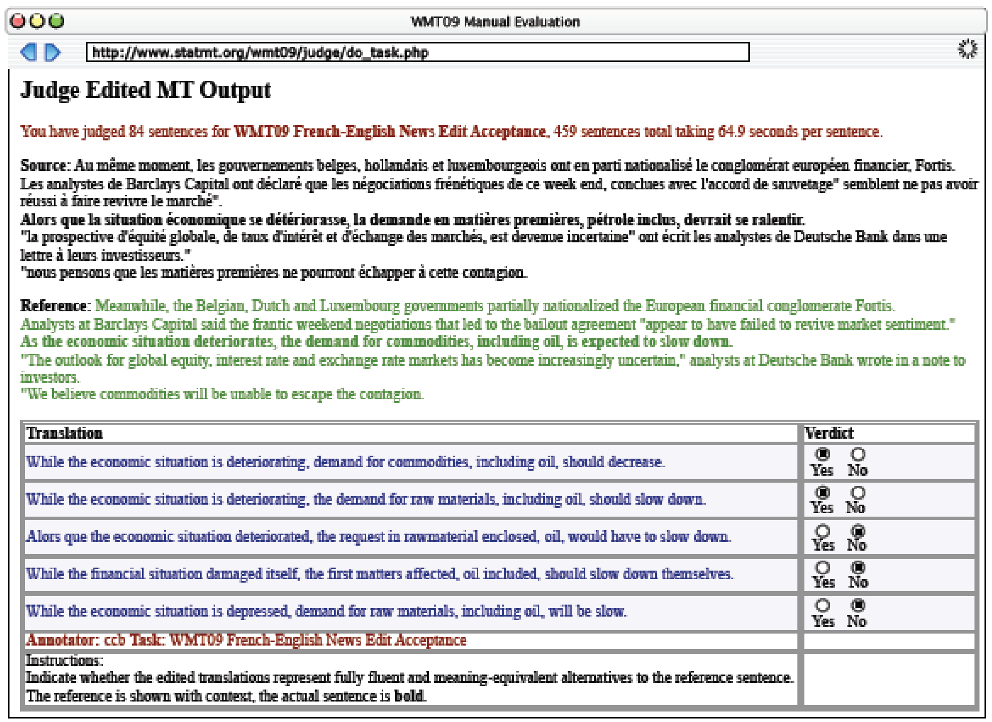


Figure 2: This screenshot shows an annotator editing the output of a machine translation system.

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.

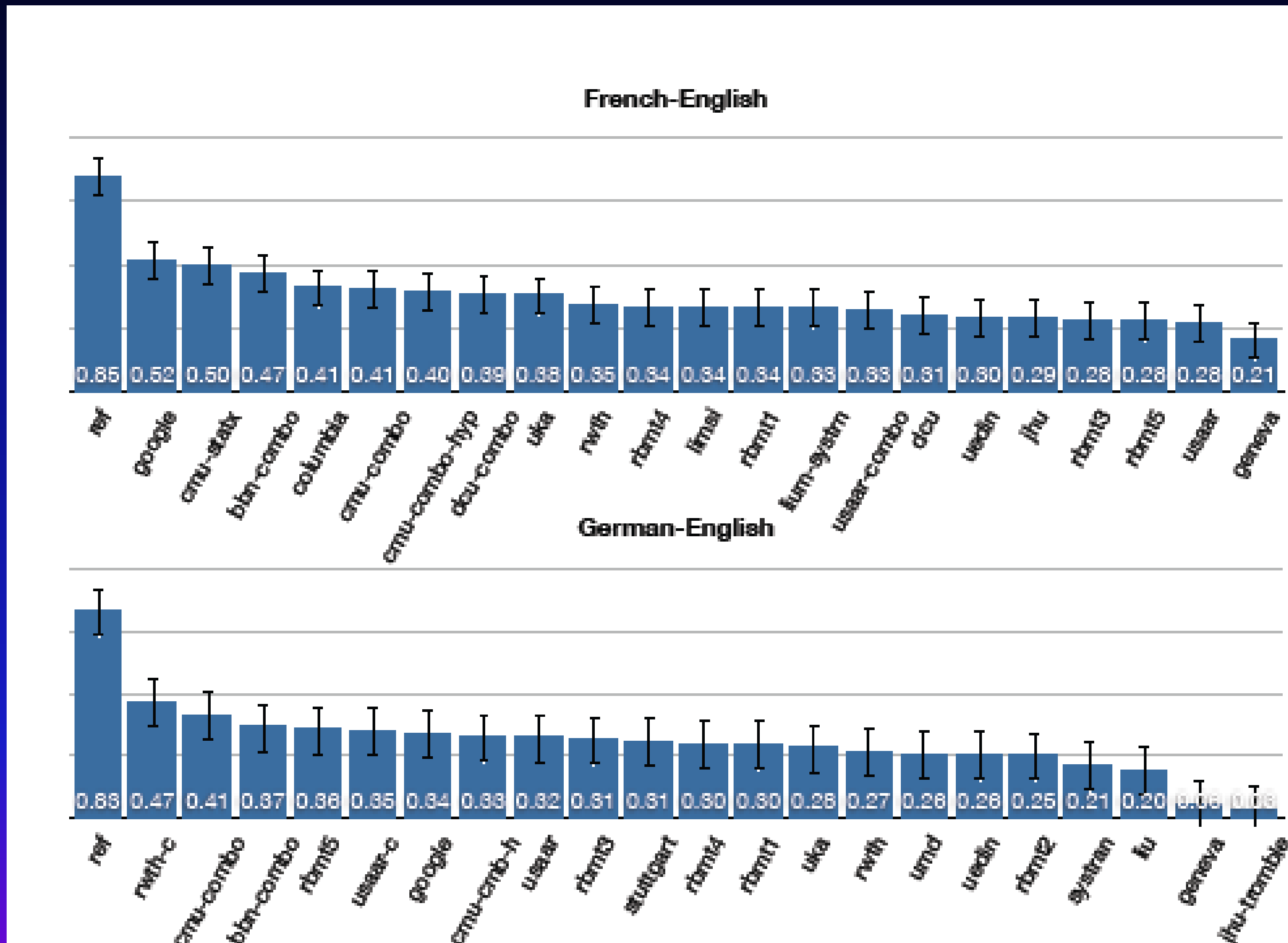
Reference: Meanwhile, the Belgian, Dutch and Luxembourg governments partially nationalized the European financial conglomerate Fortis.
Analysts at Barclays Capital said the frantic weekend negotiations that led to the bailout agreement "appear to have failed to revive market sentiment."
As the economic 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
ref
google
cmu-statx
bbn-combo
columbia
cmu-combo
cmu-combo-hyp
dcu-combo
uka
rwth
rbmt4
limsi
rbmt1
lium-systrn
usaar-combo
dcu
uedin
jhu
rbmt3
rbmt5
usaar
geneva
German-English
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

LANGUAGE PAIRS	INTER-ANNOTATOR AGREEMENT $P(A)$	$P(E)$	κ	INTRA-ANNOTATOR AGREEMENT $P(A)$	$P(E)$	κ
Czech-English	0.567	0.405	0.272	0.660	0.405	0.428
English-Czech	0.576	0.383	0.312	0.566	0.383	0.296
German-English	0.595	0.401	0.323	0.733	0.401	0.554
English-German	0.598	0.394	0.336	0.732	0.394	0.557
Spanish-English	0.540	0.408	0.222	0.792	0.408	0.648
English-Spanish	0.504	0.398	0.176	0.566	0.398	0.279
French-English	0.568	0.406	0.272	0.719	0.406	0.526
English-French	0.519	0.388	0.214	0.634	0.388	0.401
WMT12	0.568	0.396	0.284	0.671	0.396	0.455
WMT11	0.601	0.362	0.375	0.722	0.362	0.564

Table 3: Inter- and intra-annotator agreement rates for the WMT12 manual evaluation. For comparison, the WMT11 rows contain the results from the European languages individual systems task (Callison-Burch et al. (2011), Table 7).

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

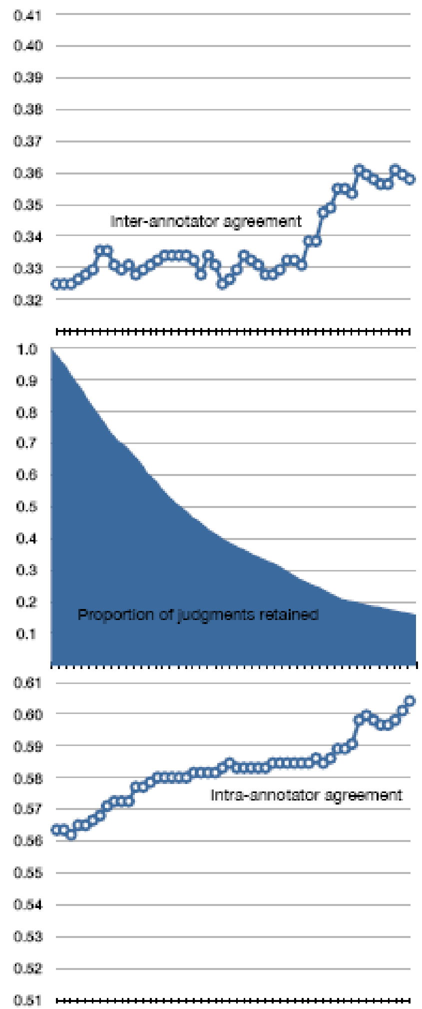


Figure 4: The effect of discarding every annotators' initial judgments, up to the first 50 items

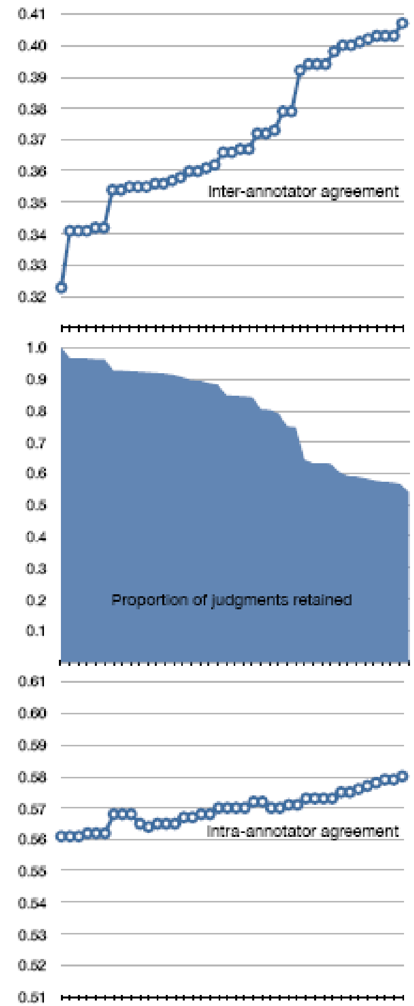


Figure 5: The effect of removing annotators with the lowest agreement, disregarding up to 40 annotators

Normalizing Human Bias

- Human judgments using absolute scales (Likert Scores) typically exhibit subjective biases among judges
- Normalizing scores across judges can significantly improve inter-coder agreement
- Several normalization methods have been proposed in recent years
- One example: (Blatz et al. 2003)
  - Normalize the scores into a continuous space [0-1] by mapping each discrete score s to the fraction of judgments of score <= s

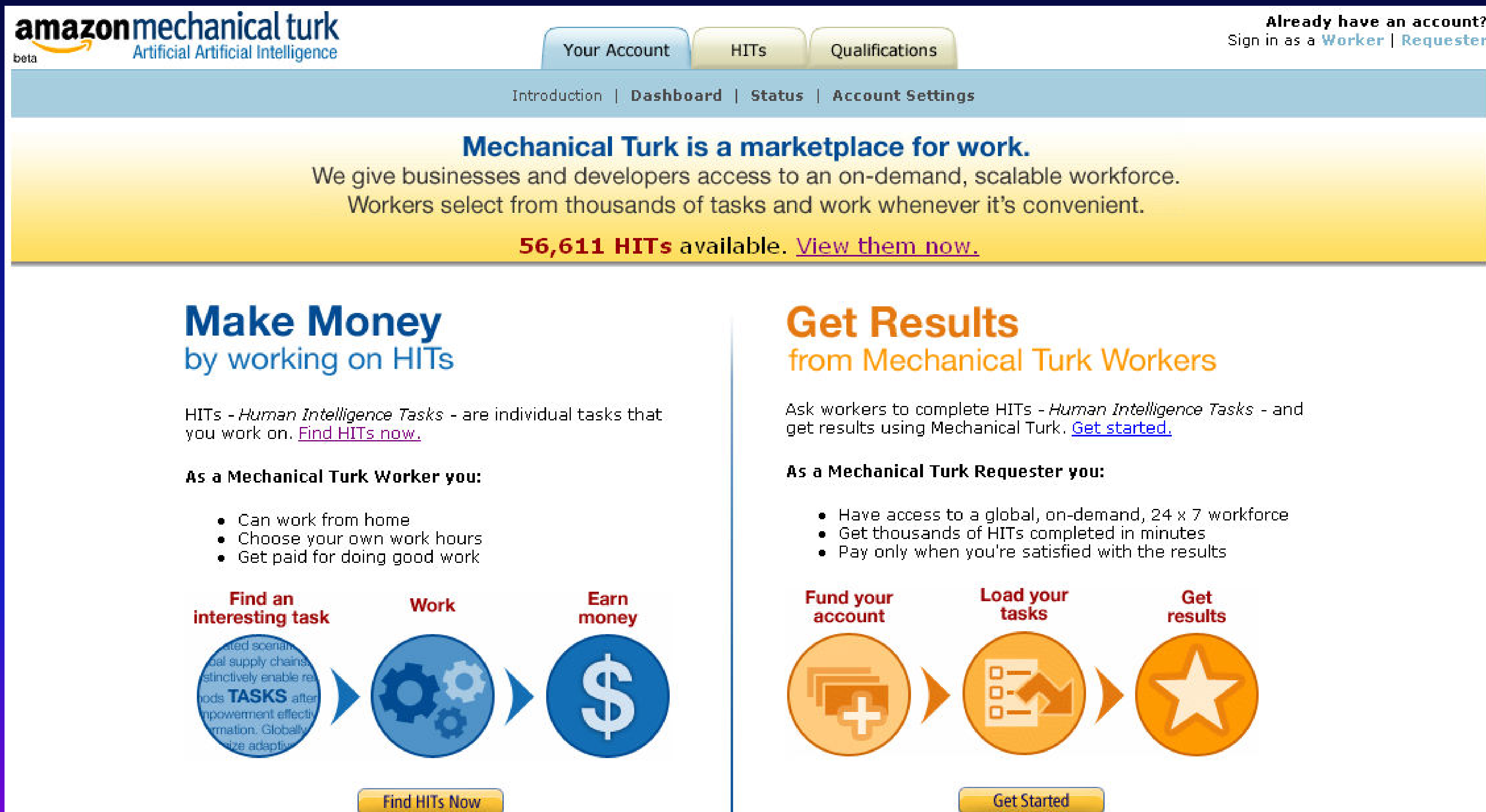
Cost and Quality Issues

- **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!

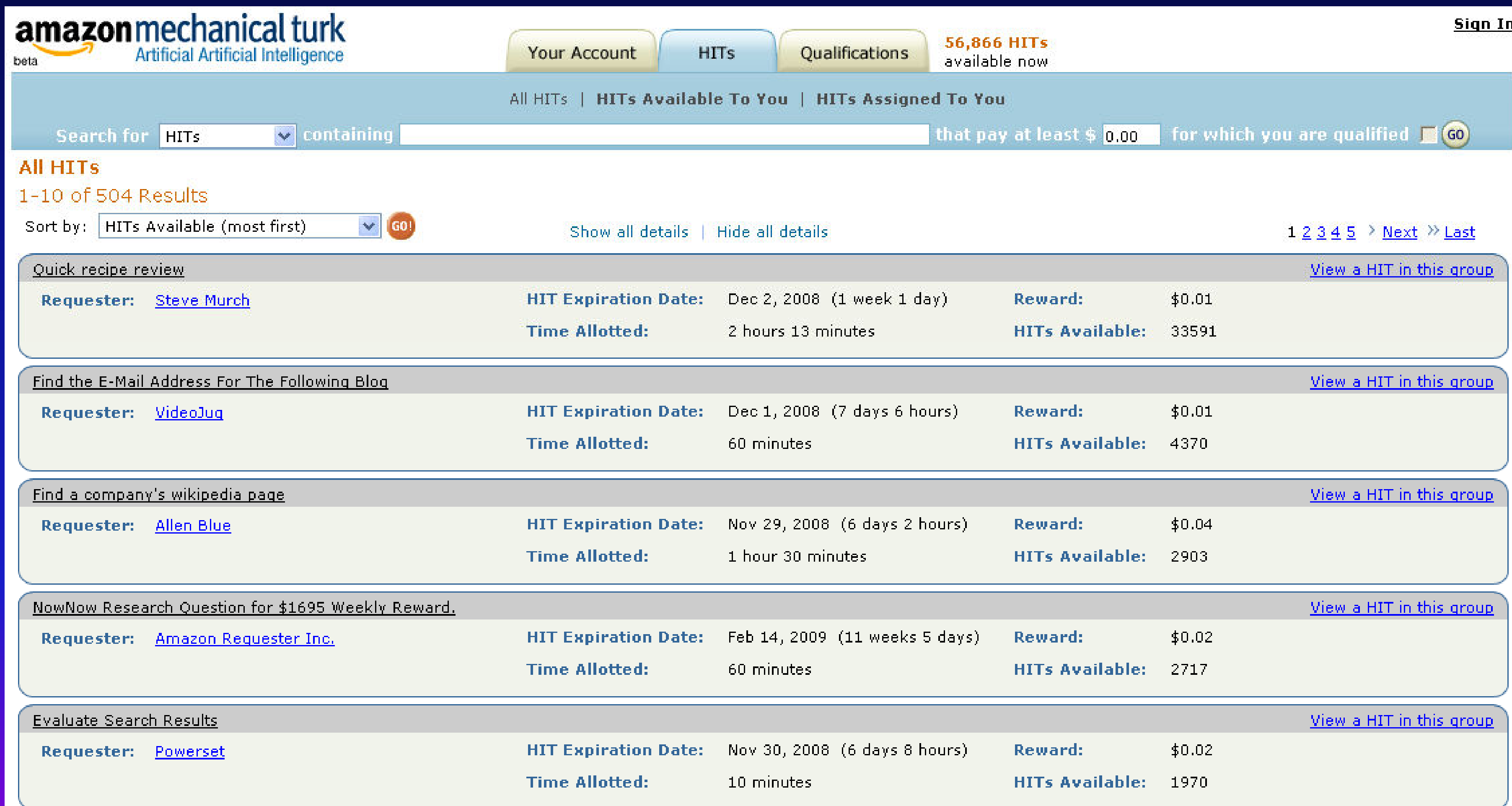
Human Evaluations Using Crowd-Sourcing

- 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 في عالم العولمة موسيفيني شدد على الحاجة إلى دول الكوميسا والدول الافريقية الى التوحد من أجل منحهم فرصة أفضل في عالم العولمة.
3 - Good (جيد):	وستبلغ القيمة المضافة للصناعة 328 مليار يوان بزيادة 12 بالمئة وقيمة الصادرات مئة مليار دولار امريكي بزيادة 8 بالمئة. في هذه الصناعة ذات القيمة المضافة ستكون 328 مليار يوان ، بزيادة 12 ٪ ، بينما ارتفعت 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 بذهب إلى; بيروت الغربية لزياد دويري

Task:

Human translation (ترجمة بشرية):	مائة فنان من 61 دولة يشاركون في اول 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 جداً)

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

Submit

Summary

- Human assessment of MT output is still extremely important… even though it is difficult to do reliably, and there is no clear consensus on best practice methods
- Human and automatic metrics are both essential in modern MT development and serve different purposes
- Good human metrics greatly help in developing good automatic metrics

Questions?