MetricsMATR

NIST is pleased to introduce the MetricsMATR Challenge, a new series of research challenge events for machine translation (MT) metrology promoting the development of innovative, even revolutionary, MT metrics. MetricsMATR focuses entirely on MT metrics.

Introduction

NIST has been conducting formal evaluations of machine translation (MT) technology since 2002, and while the evaluations have been successful, there is still a need for a better understanding of exactly how useful the state-of-the-art technology is, and how to best interpret the scores reported during evaluation.

This need exists primarily due to the shortcomings with the current methods employed for the evaluation of Machine Translation technology:

1. Automatic metrics have not yet been proved able to consistently predict the usefulness, adequacy, and reliability of MT technologies.
2. Automatic metrics have not demonstrated that they are as meaningful in target languages other than English.
3. Human assessments are expensive, slow, subjective, and are difficult to standardize. Furthermore they only pertain to the translations evaluated, and are of no use even to updated translations from the same system.
4. Both automatic metrics and human assessments need more insights into what properties of the translation should be evaluated, as well as insights into how to evaluate those properties.
5. Some MT technology approaches evaluated incorporate algorithms that optimize scores on MT metric(s). These optimizations fail in the same respects that the metrics fail.

These problems, and the need to overcome them through the development of improved automatic (and even semi-automatic) metrics, have been a constant point of discussion at past NIST MT evaluations. Without more appropriate metrics to address these shortcomings, the impact of formative and summative MT technology evaluations will remain limited.

NIST is running a new MT evaluation series "MetricsMATR" designed to address this need for improved, even revolutionary, MT metrics.

More details regarding this evaluation can be found on the MetricsMATR home page.

Data

The Metrics MATR evaluation set is composed of data from different sources. It has several language pairs, data genres, and human assessment types.

Tracks

- **Single reference track**
  For this track NIST analyzes metric performance when limiting the evaluation data to one pre-selected reference translation. We did not mix reference producers in each of the evaluation subsets, rather we choose the one that exhibited general consistency in quality.
- **Multiple references track (4 reference translations)**
  For the multiple references track, NIST analyzes metric performance when several reference translations are available to the metrics. Some portions of the evaluation set have a single reference translation; results on these data are not reported for this track. Instead, all data that have multiple reference translations (always four distinct translations) are used.

Metrics

This report includes 39 metrics, including 7 baseline metrics. More details are available on the metrics page.

Correlation Results

Our correlation analysis of MetricsMATR 2008 metrics will continue to expand for quite some time. Click here for the root node of our analysis.

References

References to this report should cite:

- Przybocki, M.; Peterson, K.; Bronsart, S.; *Official results of the NIST 2008 "Metrics for MAchine TRanslation" Challenge (MetricsMATR08)*, http://nist.gov/speech/tests/metricsmatr/2008/results/

Disclaimer

Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the (NIST), nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

Data Set

The MetricsMATR 2008 evaluation data set is not to be publicly released. Portions will be reused for future NIST MT evaluations.

Primary Evaluation Set

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

Secondary Evaluation Set

Origin	Source Language	Target Language	Genre(s)	Words (est.)	Systems
CESTA, run1	Arabic	French	General	28,000	2
	English	French	General	21,500	5
CESTA, run2	Arabic	French	Health	20,000	1
	English	French			

Metrics

Baseline Metrics

Metric Name	Affiliation	Link	Participating track(s)	
			Single Reference	Multiple References

Metric Name	Affiliation	Link	Participating track(s)	
			Single Reference	Multiple References
BLEU-1	IBM	http://talesdemo.watson.ibm.com/BLEU	yes	yes
BLEU-4	IBM	http://talesdemo.watson.ibm.com/BLEU	yes	yes
BLEU-v11b	National Institute of Standards and Technology	http://www.nist.gov/speech/tools/	yes	yes
BLEU-v12	National Institute of Standards and Technology	http://www.nist.gov/speech/tools/	yes	yes
METEOR-v0.6	Carnegie Mellon University	http://www.cs.cmu.edu/~alavie/METEOR/	yes	yes
NIST-v11b	National Institute of Standards and Technology	http://www.nist.gov/speech/tools/	yes	yes
TER-v0.7.25	University of Maryland / BBN Technologies	http://www.cs.umd.edu/~snover/tercom/	yes	yes

Site Metrics

Metric Name	Affiliation	Link	Participating track(s)	
			Single Reference	Multiple References
4-GRR	University of Southern California, Information Sciences Institute		yes	yes
ATEC1	City University of Hong Kong		yes	yes
ATEC2	City University of Hong Kong		yes	yes
ATEC3	City University of Hong Kong		yes	yes
ATEC4	City University of Hong Kong		yes	yes
BEwT-E	University of Southern California, Information Sciences Institute		yes	yes
Badger	BabbleQuest		yes	yes
BadgerLite	BabbleQuest		yes	yes
Bleu-sbp	University of Southern California, Information Sciences Institute		yes	yes
BleuSP	RWTH Aachen University		yes	yes
CDer	RWTH Aachen University		yes	yes
DP-Or	Universitat Politècnica de Catalunya, LSI		yes	yes
DP-Orp	Universitat Politècnica de Catalunya, LSI		yes	yes
DR-Or	Universitat Politècnica de Catalunya, LSI		yes	yes
EDPM	University of Washington		yes	yes
LET	Harbin Institute of Technology, School of Computer Science and Technology		yes	yes
METEOR-ranking	Carnegie Mellon University		yes	yes
MaxSim	National University of Singapore		yes	yes
Meteor-v0.7	Carnegie Mellon University		yes	yes
RTE	Stanford University		yes	no
RTE-MT	Stanford University		yes	no
SEPIA1	Columbia University		yes	yes
SEPIA2	Columbia University		yes	yes
SNR	Harbin Institute of Technology, School of Computer Science and Technology		yes	yes
SR-Or	Universitat Politècnica de Catalunya, LSI		yes	yes

Metric Name	Affiliation	Link	Participating track(s)	
			Single Reference	Multiple References
SVM-Rank	Harbin Institute of Technology, School of Computer Science and Technology		yes	yes
TERp	University of Maryland / BBN Technologies		yes	yes
ULCh	Universitat Politècnica de Catalunya, LSI		yes	yes
ULCopt	Universitat Politècnica de Catalunya, LSI		yes	yes
invWer	RWTH Aachen University		yes	yes
mBLEU	Carnegie Mellon University		yes	yes
mTER	Carnegie Mellon University		yes	yes

Correlation Results

- Correlation results for three commonly used correlation statistics are included.
- Note that columns are sortable.
- One graph is always included, regardless of the Human Assessment type:
  - "graph_scatterplot" is a traditional scatterplot. Metrics scores are on the Y axis while Human Assessment scores are on the X axis.

Human Assessment Type: Adequacy, 7-point scale, straight average

Judges were presented with a reference translation and a candidate sentence to evaluate. They answered the following "quantitative" question: "*How much of the meaning expressed in the Reference translation is also expressed in the System translation?*" on a 7-point scale ranging from 1 (None) to 7 (All).
Each segment assessed received at least two judgments from two different judges.
The segment score is the average of all (two or more) scores given on this segment. The document or system score is computed as the weighted (by segment length) average of segment scores

Additional charts:

- "graph_category": A sorted scatter plot. All scores for corresponding assessment categories are binned and re-sorted.
- "graph_category_2": A box-and-whisker type of graph, using the same bins as the previous chart.

Target Language: English

- Segment-level correlation
- Document-level correlation
- System-level correlation

Human Assessment Type: Adequacy, Yes-No qualitative question, proportion of Yes assigned

The 7-point scale adequacy question described above was followed by a second, more "qualitative" question: "*Does the Machine translation mean essentially the same as the Reference translation?*". Judges did not have to answer this binary Yes/No question if their answer to the preceding question was 4 (Half) or less, in which case the answer was considered to be 'No' by default.
Each segment assessed received at least two judgments from two different judges.
The score is the number of 'Yes' assigned divided by the total number of judgments. The proportion is computed identically for segment, document and system level scores.

Additional charts:

- "graph_YN" (segment-level only): Shows the segments that received 'No' on the left side of the graph, and segments that received 'Yes' on the right side of the graph. Segments are not ordered. The Y axis value is the metric score.
- "graph_ILR" (document-level only): Shows several graphs, one per ILR level. All segments that belong to one of the documents that share the same ILR are presented on the same graph. The X axis represents the metric score. Segments are ordered by metric score. Each graph has three series: segments that received only 'Yes' judgments are shown in blue, segments that receive only 'No' judgments are shown in red, and segments that received a mix of 'Yes' and 'No' are shown in yellow.

Target Language: English

- Segment-level correlation
- Document-level correlation
- System-level correlation

Human Assessment Type: Preferences, Pair-wise comparison across systems

Two candidate translations of the same segment from two different systems were presented to a judge, along with a reference translation. The judge decided which candidate translation he/she prefers, with 'No preference' available as a third choice.
A full pair-wise comparison across systems was performed on a selected number of segments.
The segment score represents the number of times the given system segment was preferred, divided by the number of judgments involving this same system segment. The proportion is computed identically for segment, document, or system-level scores.

Target Language: English

- Segment-level correlation
- Document-level correlation
- System-level correlation

Human Assessment Type: Adjusted Probability that a Concept is Correct

Source text low level concepts were identified beforehand. Several bilingual judges (5 for Farsi, 6 for Arabic) then looked for these concepts in the candidate sentences, comparing it against the annotated source sentence and marking deletions, substitutions, and insertions.
A segment score is the number of correctly conveyed concepts, divided by the total number of concepts identified in the source sentence (including concepts identified by the judges as inserted concepts). Measures are aggregated for document and system scores.

Target Language: English

- Segment-level correlation
- Document-level correlation
- System-level correlation

Human Assessment Type: Adequacy, 4-point scale

A set of bilingual judges (5 for Farsi, 6 for Arabic) graded the adequacy of a candidate translation by comparing it to the source sentence in a two-step process, first identifying the candidate as more adequate or more inadequate, then within one of those, as completely adequate or tending adequate, or as inadequate or tending inadequate. Thus possible scores range from 1 (Inadequate) to 4 (Completely adequate).
A segment score is the average of all scores given on this segment. A document or system score is the weighted (by segment length) average of segment scores.

Target Language: English

- Segment-level correlation
- Document-level correlation
- System-level correlation

Human Assessment Type: Adequacy, 5-point scale

Judges evaluated the adequacy of Arabic-to-French and English-to-French translations. Each segment received one judgment. Judgments were performed using a 5-point scale. Document and system level scores are the weighted (by segment length) averages of segment scores.

Target Language: French

- Segment-level correlation
- Document-level correlation
- System-level correlation

Human Assessment Type: Fluency, 5-point scale

Judges evaluated the fluency of Arabic-to-French and English-to-French translations. Each segment received one judgment. Judgments were performed using a 5-point scale. Document and system level scores are the weighted (by segment length) averages of segment scores.

Target Language: French

- Segment-level correlation
- Document-level correlation
- System-level correlation

Human Assessment Type: HTER

A human annotator modifies a candidate translation so that it has the same meaning as a reference translation. Emphasis is on as few edits as possible to achieve the same meaning. Then, this modified text is used as a single reference to compute the TER score for the candidate sentence. Document and system level scores are computed as weighted (by segment length) averages of segment scores.

Target Language: English

- Segment-level correlation
- Document-level correlation
- System-level correlation