Choosing the Right Evaluation for Machine Translation

An Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks

Michael Denkowski and Alon Lavie

Language Technologies Institute
Carnegie Mellon University

November 3, 2010

Introduction

How do we evaluate performance of machine translation systems?

Introduction

How do we evaluate performance of machine translation systems?

- Simple: have humans evaluate translation quality

Introduction

How do we evaluate performance of machine translation systems?

- Simple: have humans evaluate translation quality

Not so simple:

- Can this task be completed reliably?
- Can judgments be collected efficiently?
- What types of judgments are most informative?
- Are judgments usable for developing automatic metrics?

Related Work

ACL Workshop for Statistical Machine Translation (WMT) [Callison-Burch et al.2007]

- Compares absolute and relative judgment tasks, metric performance on tasks

Related Work

ACL Workshop for Statistical Machine Translation (WMT) [Callison-Burch et al.2007]

- Compares absolute and relative judgment tasks, metric performance on tasks

NIST Metrics for Machine Translation Challenge (MetricsMATR) [Przybocki et al.2008]

- Compare metric performance on various tasks

Related Work

ACL Workshop for Statistical Machine Translation (WMT) [Callison-Burch et al.2007]

- Compares absolute and relative judgment tasks, metric performance on tasks

NIST Metrics for Machine Translation Challenge (MetricsMATR) [Przybocki et al.2008]

- Compare metric performance on various tasks

Snover et al. (TER-plus) [Snover et al.2009]

- Tune TERp to adequacy, fluency, HTER judgments, compare parameters and correlation

This Work

Deeper exploration of judgment tasks

- Motivation, design, practical results
- Challenges for human evaluators

This Work

Deeper exploration of judgment tasks

- Motivation, design, practical results
- Challenges for human evaluators

Examine behavior of tasks by tuning versions of the METEOR-NEXT metric

- Fit metric parameters for multiple tasks and years
- Examine parameters, correlation with human judgments
- Determine task stability, reliability

Adequacy

Introduced by Linguistics Data Consortium for MT evaluation [LDC2005]

Adequacy: how much meaning expressed in reference is expressed in MT translation hypothesis?

5: All **4: Most** **3: Much** **2: Little** **1: None**

Fluency: how well-formed is hypothesis in target language?

5: Flawless **4: Good** **3: Non-native**

2: Disfluent **1: Incomprehensible**

Adequacy

Two scales better than one?

- High correlation between adequacy and fluency (WMT 2007)
- NIST Open MT [Przybocki2008]: adequacy only, 7 point scale (precision vs accuracy)

Problems encountered:

- Low inter-annotator agreement: $K = 0.22$ for adequacy
  $K = 0.25$ for fluency
- Severity of error: how to penalize single term negation?
- Difficulty with boundary cases (3 or 4?)

Adequacy

Good news:

- Multiple annotators help: scores averaged or otherwise normalized
- Consensus among judges approximates actual adequacy
- Clear objective function for metric tuning: segment-level correlation with normalized adequacy scores

Ranking

Directions: simply rank multiple translations from best to worst.

- Avoid difficulty of absolute judgment, use relative comparison
- Allow fine-grained judgments of translations in same adequacy bin
- Facilitated by system outputs from WMT evaluations

Ranking

Motivation:

Inter-Annotator Agreement			
Judgment Task	$P(A)$	$P(E)$	K
Adequacy	0.38	0.20	0.23
Fluency	0.40	0.20	0.25
Ranking	0.58	0.33	0.37

Intra-Annotator Agreement			
Judgment Task	$P(A)$	$P(E)$	K
Adequacy	0.57	0.20	0.47
Fluency	0.63	0.20	0.54
Ranking	0.75	0.33	0.62

Table: Annotator agreement for absolute and relative judgment tasks in WMT07

Ranking

Complication: tens of similar systems (WMT09, WMT10)

Ranking

Complication: tens of similar systems (WMT09, WMT10)

Task: Spanish-to-English

Reference: Discussions resumed on Friday.

Ranking

Complication: tens of similar systems (WMT09, WMT10)

Task: Spanish-to-English

Reference:	Discussions resumed on Friday.
System 1:	Discussions resumed on Monday.

Ranking

Complication: tens of similar systems (WMT09, WMT10)

Task: Spanish-to-English

Reference:	Discussions resumed on Friday.
System 1:	Discussions resumed on Monday.
System 2:	Discussions resumed on .

Ranking

Complication: tens of similar systems (WMT09, WMT10)

Task: Spanish-to-English

Reference:	Discussions resumed on Friday.
System 1:	Discussions resumed on Monday.
System 2:	Discussions resumed on .
System 3:	Discussions resumed on Viernes.

What is the correct ranking for these translations?

Ranking

Even worse: common case in WMT10 evaluation

Reference:	p_1 p_2 p_3 p_4
System 1:	p_1 incorrect
System 2:	p_2 incorrect, p_2 half length of p_1
System 3:	p_3 and p_4 incorrect, combined length $<$ p_1 or p_2
System 4:	Content words correct, function words missing
System 5:	Main verb incorrectly negated

Clearly different classes of errors present - all ties?

Ranking

Overall complications:

- Different numbers of difficult-to-compare errors
- Judges must keep multiple long sentences in mind
- All ties? Universal confusion inflates annotator agreement

Bad news:

- Multiple annotators can invalidate one another
- Normalize with ties? Ties must be discarded when tuning metrics.

Post-Editing

Motivation: eliminate need for absolute or relative judgments

- Judges correct MT output - no scoring required
- Automatic measure (TER) determines cost of edits
- HTER widely adopted by GALE project [Olive2005]

Post-Editing

Challenges:

- Accuracy of scores limited by automatic measure (TER)
- Inserted function word vs inserted negation term?
- Need for reliable, accurate, automatic metrics

Good news:

- Multiple annotators help: approach true minimum for edits
- Byproducts: set of edits, additional references
- Segment level scores allow simple metric tuning

Metric Tuning

Experiment: Use METEOR-NEXT to explore human judgment tasks

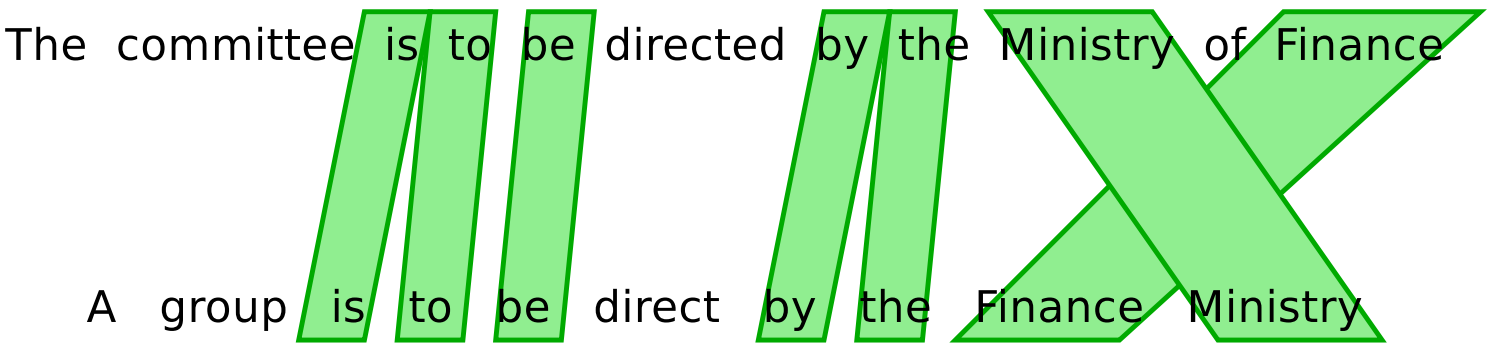
- Tune versions of METEOR-NEXT on each type of judgment
- Examine parameters and correlation across tasks, evaluations
- Determine which judgment tasks are most stable
- Evaluate performance of METEOR-NEXT on tasks

METEOR-NEXT Scoring

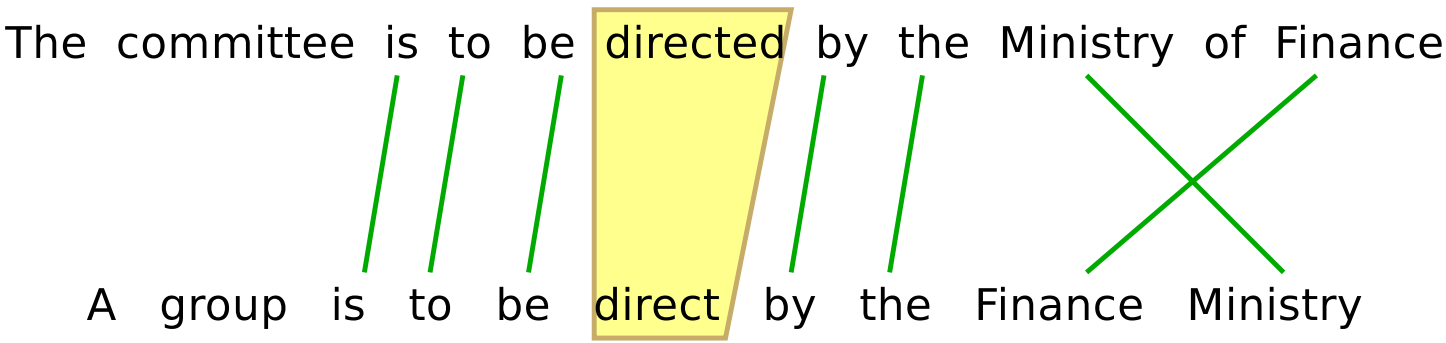
The committee is to be directed by the Ministry of Finance

A group is to be direct by the Finance Ministry

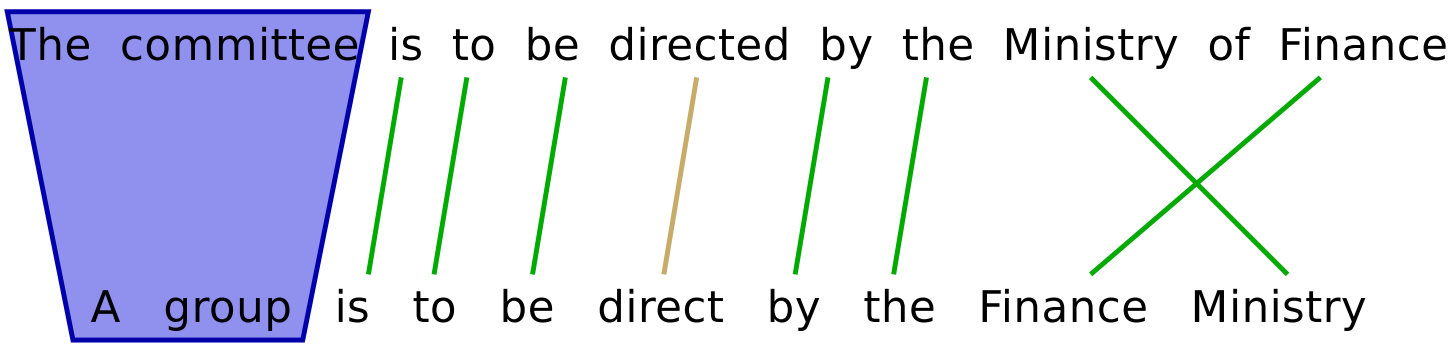
METEOR-NEXT Scoring



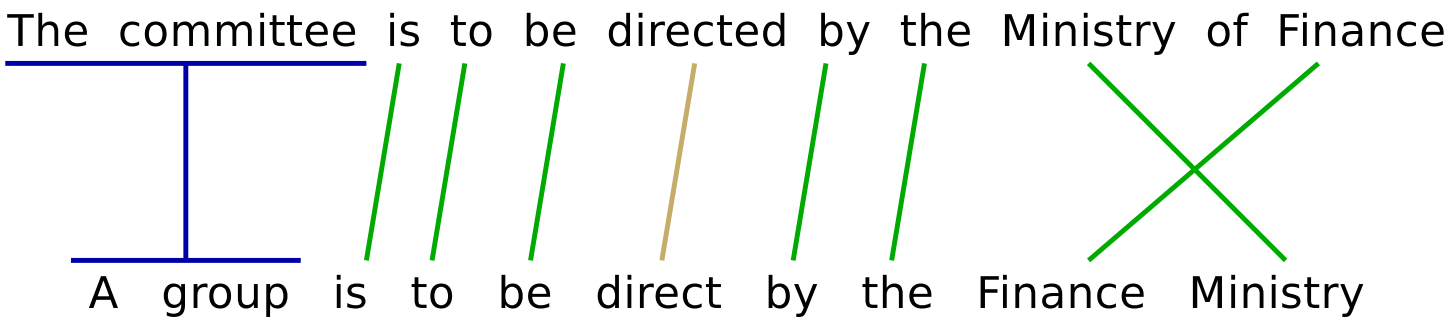
METEOR-NEXT Scoring



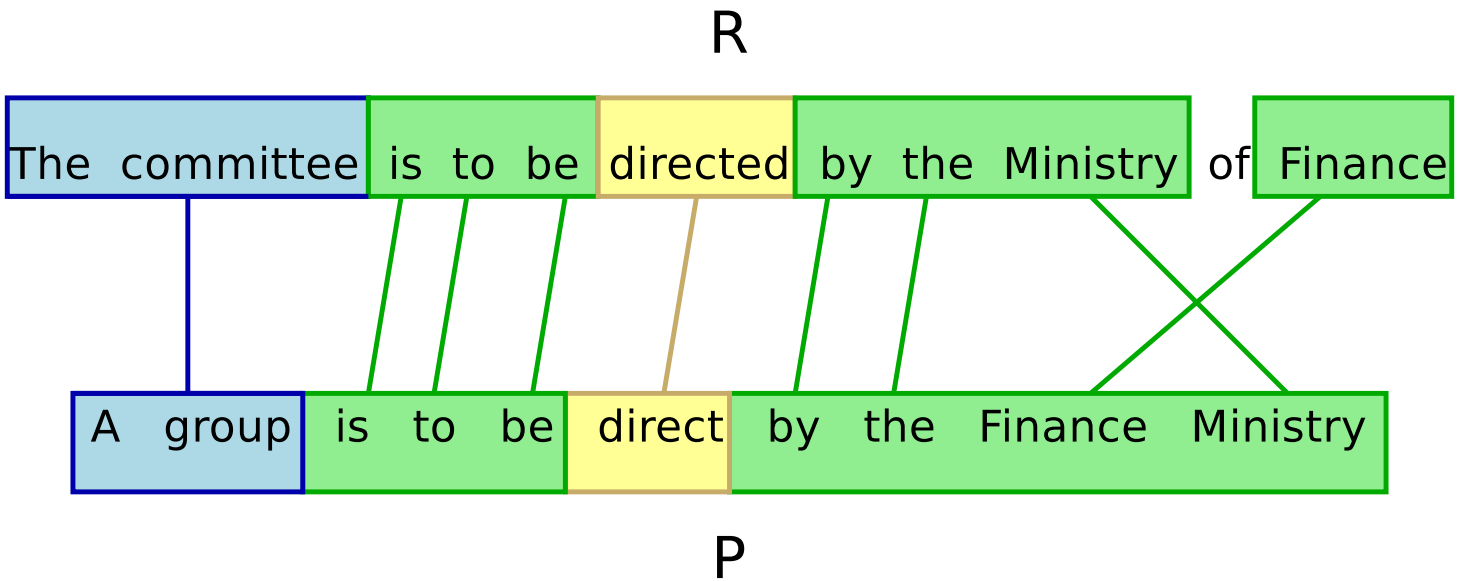
METEOR-NEXT Scoring



METEOR-NEXT Scoring

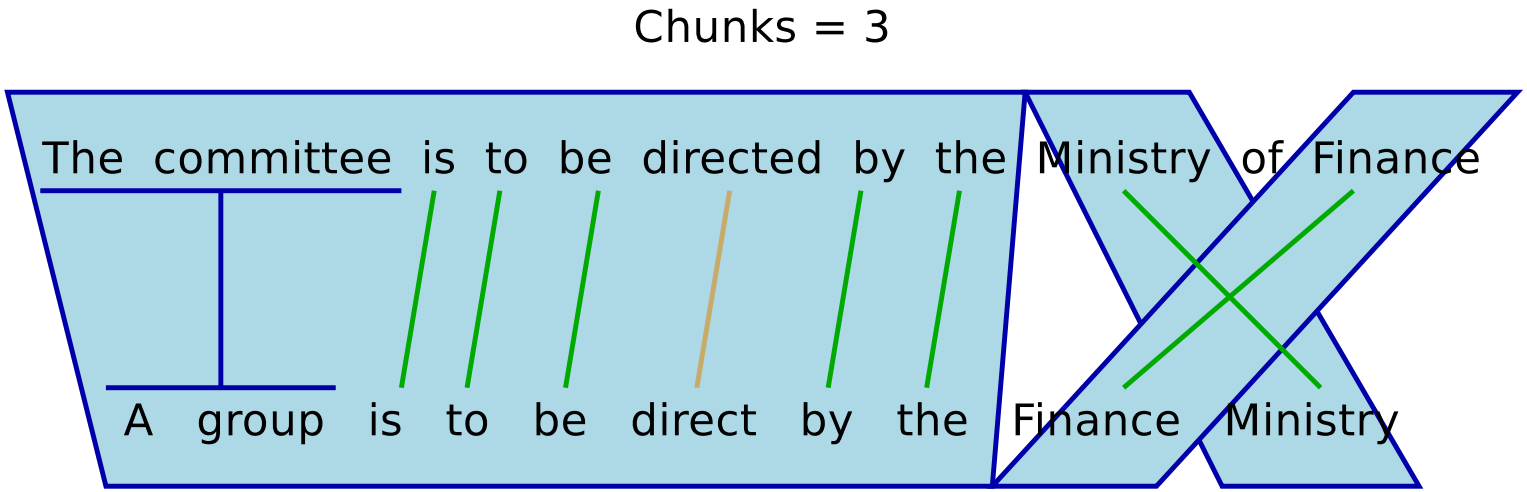


METEOR-NEXT Scoring



Matches weighted by type: $m_{exact} + m_{stem} + m_{par}$

METEOR-NEXT Scoring



Chunk: contiguous, ordered matches

METEOR-NEXT Scoring

$$Score = \left(1 - \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}\right) \cdot F_{mean} \qquad F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R}$$

METEOR-NEXT Scoring

$$Score = \left(1 - \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}\right) \cdot F_{mean} \qquad F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R}$$

- α – Balance between P and R
- β, γ – Control severity of fragmentation penalty
- w_{stem} – Weight of stem match
- w_{syn} – Weight of WordNet synonym match
- w_{par} – Weight of paraphrase match

METEOR-NEXT Scoring

$$Score = \left(1 - \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}\right) \cdot F_{mean} \qquad F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R}$$

- α – Balance between P and R
- β, γ – Control severity of fragmentation penalty
- w_{stem} – Weight of stem match
- w_{syn} – Weight of WordNet synonym match
- w_{par} – Weight of paraphrase match

METEOR-NEXT Scoring

$$Score = \left(1 - \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}\right) \cdot F_{mean} \qquad F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R}$$

- α – Balance between P and R
- β, γ – Control severity of fragmentation penalty
- w_{stem} – Weight of stem match
- w_{syn} – Weight of WordNet synonym match
- w_{par} – Weight of paraphrase match

METEOR-NEXT Tuning

Tuning versions of METEOR-NEXT

- Align all hypothesis/reference pairs once
- Optimize parameters using grid search
- Select objective function appropriate for task

Metric Tuning Results

Parameter stability for judgment tasks:

Metric Tuning Results

Parameter stability for judgment tasks:

Tuning Data		α	β	γ	w_{stem}	w_{syn}	w_{para}
MT08	Adequacy	0.60	1.40	0.60	1.00	0.60	0.80
MT09	Adequacy	0.80	1.10	0.45	1.00	0.60	0.80

Metric Tuning Results

Parameter stability for judgment tasks:

Tuning Data		α	β	γ	w_{stem}	w_{syn}	w_{para}
MT08	Adequacy	0.60	1.40	0.60	1.00	0.60	0.80
MT09	Adequacy	0.80	1.10	0.45	1.00	0.60	0.80
WMT08	Ranking	0.95	0.90	0.45	0.60	0.80	0.60
WMT09	Ranking	0.75	0.60	0.35	0.80	0.80	0.60

Metric Tuning Results

Parameter stability for judgment tasks:

Tuning Data		α	β	γ	w_{stem}	w_{syn}	w_{para}
MT08	Adequacy	0.60	1.40	0.60	1.00	0.60	0.80
MT09	Adequacy	0.80	1.10	0.45	1.00	0.60	0.80
WMT08	Ranking	0.95	0.90	0.45	0.60	0.80	0.60
WMT09	Ranking	0.75	0.60	0.35	0.80	0.80	0.60
GALE-P2	HTER	0.65	1.70	0.55	0.20	0.60	0.80
GALE-P3	HTER	0.60	1.70	0.35	0.20	0.40	0.80

Metric Tuning Results

Metric correlation for judgment tasks:

Tuning Best

Metric Tuning Results

Metric correlation for judgment tasks:

Tuning Best

		Adequacy (r)		Ranking (consist)		HTER (r)	
Metric	Tuning	MT08	MT09	WMT08	WMT09	G-P2	G-P3
BLEU	N/A	0.504	0.533	–	0.510	-0.545	-0.489
TER	N/A	-0.439	-0.516	–	0.450	0.592	0.515
METEOR	N/A	0.588	0.597	0.512	0.490	-0.625	-0.568

Metric Tuning Results

Metric correlation for judgment tasks:

Tuning Best

		Adequacy (r)		Ranking (consist)		HTER (r)	
Metric	Tuning	MT08	MT09	WMT08	WMT09	G-P2	G-P3
BLEU	N/A	0.504	0.533	–	0.510	-0.545	-0.489
TER	N/A	-0.439	-0.516	–	0.450	0.592	0.515
METEOR	N/A	0.588	0.597	0.512	0.490	-0.625	-0.568
M-NEXT	MT08	0.620	0.625	0.630	0.614	-0.638	-0.590
M-NEXT	MT09	0.612	0.630	0.637	0.617	-0.636	-0.589

Metric Tuning Results

Metric correlation for judgment tasks:

Tuning Best

		Adequacy (r)		Ranking (consist)		HTER (r)	
Metric	Tuning	MT08	MT09	WMT08	WMT09	G-P2	G-P3
BLEU	N/A	0.504	0.533	–	0.510	-0.545	-0.489
TER	N/A	-0.439	-0.516	–	0.450	0.592	0.515
METEOR	N/A	0.588	0.597	0.512	0.490	-0.625	-0.568
M-NEXT	MT08	0.620	0.625	0.630	0.614	-0.638	-0.590
M-NEXT	MT09	0.612	0.630	0.637	0.617	-0.636	-0.589
M-NEXT	WMT08	0.598	0.626	0.643	0.621	-0.629	-0.573
M-NEXT	WMT09	0.601	0.624	0.635	0.629	-0.628	-0.578

Metric Tuning Results

Metric correlation for judgment tasks:

Tuning Best

		Adequacy (r)		Ranking (consist)		HTER (r)	
Metric	Tuning	MT08	MT09	WMT08	WMT09	G-P2	G-P3
BLEU	N/A	0.504	0.533	–	0.510	-0.545	-0.489
TER	N/A	-0.439	-0.516	–	0.450	0.592	0.515
METEOR	N/A	0.588	0.597	0.512	0.490	-0.625	-0.568
M-NEXT	MT08	0.620	0.625	0.630	0.614	-0.638	-0.590
M-NEXT	MT09	0.612	0.630	0.637	0.617	-0.636	-0.589
M-NEXT	WMT08	0.598	0.626	0.643	0.621	-0.629	-0.573
M-NEXT	WMT09	0.601	0.624	0.635	0.629	-0.628	-0.578
M-NEXT	G-P2	0.616	0.623	0.632	0.615	-0.640	-0.596
M-NEXT	G-P3	0.610	0.618	0.636	0.617	-0.638	-0.600

Metric Tuning Results

Metric correlation for judgment tasks:

Tuning Best Test Best

		Adequacy (r)		Ranking (consist)		HTER (r)	
Metric	Tuning	MT08	MT09	WMT08	WMT09	G-P2	G-P3
BLEU	N/A	0.504	0.533	–	0.510	-0.545	-0.489
TER	N/A	-0.439	-0.516	–	0.450	0.592	0.515
METEOR	N/A	0.588	0.597	0.512	0.490	-0.625	-0.568
M-NEXT	MT08	0.620	0.625	0.630	0.614	-0.638	-0.590
M-NEXT	MT09	0.612	0.630	0.637	0.617	-0.636	-0.589
M-NEXT	WMT08	0.598	0.626	0.643	0.621	-0.629	-0.573
M-NEXT	WMT09	0.601	0.624	0.635	0.629	-0.628	-0.578
M-NEXT	G-P2	0.616	0.623	0.632	0.615	-0.640	-0.596
M-NEXT	G-P3	0.610	0.618	0.636	0.617	-0.638	-0.600

Conclusions

Summary:

- Evaluation tasks have different strengths/weaknesses.
- Minimize annotator confusion / maximize impact of human evaluation

Tuning Results:

- HTER parameters generally stable, ranking and adequacy parameters fluctuate
- METEOR-NEXT tuned to HTER data has most consistent performance
- On larger scale, METEOR-NEXT is stable across tasks, evaluations

Choosing the Right Evaluation for Machine Translation

An Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks

Michael Denkowski and Alon Lavie

Language Technologies Institute
Carnegie Mellon University

November 3, 2010


Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder.
2007.
(Meta-) Evaluation of Machine Translation.
In *Proc. Second Workshop on Statistical Machine Translation*, pages 136–158.


LDC.
2005.
Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations. Revision 1.5.


Joseph Olive.
2005.
Global Autonomous Language Exploitation (GALE).
DARPA/IPTO Proposer Information Pamphlet.


M. Przybocki, K. Peterson, and S Bronsart.
2008.

Official results of the NIST 2008 "Metrics for MAchine TRanslation" Challenge (MetricsMATR08).


Mark Przybocki.
2008.
NIST Open Machine Translation 2008 Evaluation.
http://www.itl.nist.gov/iad/mig/tests/mt/2008/.


Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz.
2009.
Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric.
In *Proc. of WMT09*.