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
Pittsburgh, PA 15217
{mdenkows,alavie}@cs.cmu.edu

Abstract

This paper examines the motivation, design, and practical results of several types of human evaluation tasks for machine translation. In addition to considering annotator performance and task informativeness over multiple evaluations, we explore the practicality of tuning automatic evaluation metrics to each judgment type in a comprehensive experiment using the METEOR-NEXT metric. We present results showing clear advantages of tuning to certain types of judgments and discuss causes of inconsistency when tuning to various judgment data, as well as sources of difficulty in the human evaluation tasks themselves.

1 Introduction

The need for efficient, reliable human evaluation of machine translation (MT) output has led to the creation of several judgment tasks. Evaluations investigating the objective quality of machine translation often elicit absolute quality judgments such as adequacy or fluency ratings. Several problems are quickly encountered with this approach: annotators have difficulty agreeing on what factors constitute "good" or "bad" translations and are often unable to reproduce their own absolute scores of translation quality. Relative judgment tasks such as translation ranking address these problems by eliminating notions of objective "goodness" or "badness" of translations in favor of simpler comparisons. While these tasks show greater agreement between judges, ranking can prove difficult and confusing when translation hypotheses are nearly identical or contain difficult-to-compare errors. Post-editing tasks remove human scoring entirely, asking annotators to "fix" MT output and relying on automatic measures to determine scores based on edit data. While post-editing tasks avoid issues inherent to absolute and relative judgment tasks, they are limited by the quality of the automatic measures used. As each task has relative strengths and weaknesses, it is advantageous to determine both the sort of information that can be gleaned from collected judgments and how reliable this information will be when selecting an evaluation task.

This work examines several types of human judgment tasks across multiple evaluations. We discuss the motivation, design, and results of these tasks in both theory and practice, focusing on sources of difficulty for annotators, informativeness of results, and consistency of evaluation conditions. As it is also advantageous to develop automatic evaluation metrics to stand in for human judgments, we conduct a comprehensive experiment tuning versions of the METEOR-NEXT metric (Denkowski and Lavie, 2010) on multiple types of human judgments from multiple evaluations to determine which types of judgments are best suited for metric development.

2 Related Work

The Association for Computational Linguistics (ACL) Workshop on Statistical Machine Translation (WMT) has conducted yearly evaluations of machine translation quality as well as *meta*-evaluation of human judgments of translation quality and automatic evaluation metric performance. WMT07 (Callison-Burch et al., 2007) compares multiple

types of human MT evaluation tasks, including adequacy-fluency scale judgments and ranking judgments, across various criteria. Ranking judgments, in which annotators rank translation hypotheses of the same source sentence from different MT systems, are shown to have higher inter-annotator and intra-annotator agreement than relative and absolute adequacy-fluency judgments. The workshop also evaluates the correlation of several automatic evaluation metrics with both types of human judgments, showing that different metrics perform best on different tasks. While our work also discusses many of these points, we consider a more diverse set of judgment tasks and conduct a more in-depth analysis of the advantages and disadvantages of each task. We also specifically address the task of tuning automatic metrics to have high correlation with these types of judgments.

The 2008 NIST Metrics for Machine Translation Challenge (MetricsMATR) (Przybocki et al., 2008) comprehensively evaluates the correlation of 39 automatic MT evaluation metrics with several types of human judgments. The results indicate that metrics perform differently on different judgment tasks, supporting the notion that metrics can be designed or tuned to have improved correlation with various types of human judgments. While many further analyses can be conducted on the resulting data, the evaluation results do not directly discuss the relative merits of the included human judgment scenarios or the task of metric tuning.

Snover et al. (2009) explore several types of human judgments using TER-Plus (TERp), a highly configurable automatic MT evaluation metric. The authors tune versions of TERp to maximize correlation with adequacy, fluency, and HTER (Snover et al., 2006) scores and present an analysis of the resulting parameter values for each task. Adequacy and Fluency parameters favor recall, having low edit costs for inserting additional words in translation hypotheses and high edit costs for removing words in hypotheses, while HTER parameters are more balanced between precision and recall. In all cases, correlation with human judgments is significantly improved by tuning TERp on similar data. Our work includes a similar metric tuning experiment using the METEOR-NEXT (Denkowski and Lavie, 2010) metric on similar adequacy and HTER judgments as well as ranking judgments. We explore the metric tuning task further by comparing the performance of metric versions tuned *across* multiple evaluations and types of human judgments.

3 Human Evaluation of Machine Translation Quality

As machine translation systems aim to replicate the results of human translation, it is desirable to incorporate human judgments of translation quality into system development. However, such judgments are often costly and time-consuming to collect, motivating both the design of highly efficient, effective human evaluation tasks and development of automatic evaluation metrics to stand in for such human judgments when necessary. With these goals in mind, evaluation tasks can be examined based on performance of annotators, informativeness of resulting data, and feasibility of tuning automatic metrics to have high correlation with collected judgments.

3.1 Adequacy Judgments

Originally introduced by the Linguistics Data Consortium for evaluation of machine translation, the adequacy scale task (LDC, 2005) elicits absolute quality judgments of MT output from human annotators using straightforward numerical ranges. These judgments are traditionally split into two categories: *adequacy* and *fluency*.

Adequacy judgments ask annotators to rate the amount of meaning expressed in a reference translation that is also expressed in a translation hypothesis using following scale:
5: All
4: Most
3: Much
2: Little
1: None

Fluency judgments ask annotators to rate the well-formedness of a translation hypothesis in the target language, regardless of sentence meaning. Fluency follows the scale:
5: Flawless
4: Good
3: Non-native
2: Disfluent
1: Incomprehensible

While these scales are originally separate, the 2007 ACL Workshop on Statistical Machine Translation (Callison-Burch et al., 2007) reports high correlation between annotators' adequacy and fluency scores. In practice, annotators have difficulty drawing any meaning from highly disfluent translations, leading them to provide low adequacy scores. Similarly, for a translation to fully express the meaning of a reference, it must also be fully, or near fully fluent, as slight changes in word order and morphology can dramatically alter meaning in many languages. In addition, the separation of adequacy and fluency leads to the problem of recombining the two scores in some meaningful way for tasks such as tuning automatic metrics. The NIST Open Machine Translation Evaluation (Przybocki, 2008; Przybocki, 2009), elicits judgments of adequacy only and expands the task to use a 7-point scale, allowing for more fine-grained distinctions.

Annotator agreement for adequacy tasks can be measured using the kappa coefficient:

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

where $P(A)$ is the proportion of times annotators agree and $P(E)$ is the proportion of times annotators are expected to agree by chance. The WMT07 human evaluation reports inter-annotator kappa of 0.22 for adequacy and 0.25 for fluency, and intra-annotator kappa of 0.47 for adequacy and 0.54 for fluency (also listed in Table 1). These relatively low values illustrate the difficulty encountered when annotators have different notions of meaning expressiveness. For example, a single negation term can reverse the meaning of a long sentence. Should the annotator mark "most" of the meaning expressed as only one term is incorrect, or mark "none" of the meaning expressed since the sentence means the inverse of the reference? Annotators also disagree on which parts of sentences are most important; which phrases in a 40–50 word sentence are most vital to correct meaning? Annotators also have difficulty agreeing with themselves, as very long sentences are difficult to classify correctly and even short sentences pose difficulty when they fall on boundaries between adequacy categories.

The addition of multiple annotators effectively addresses many issues inherent to the adequacy task,

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

Inter-Annotator Agreement			
Evaluation	$P(A)$	$P(E)$	K
WMT07	0.58	0.33	0.37
WMT08	0.58	0.33	0.37
WMT09	0.55	0.33	0.32
Intra-Annotator Agreement			
Evaluation	$P(A)$	$P(E)$	K
WMT07	0.75	0.33	0.62
WMT08	0.69	0.33	0.54
WMT09	0.72	0.33	0.56

Table 2: Annotator agreement for ranking task across multiple years

as numerical judgments can be averaged or otherwise normalized (Blatz et al., 2003). This makes the task of tuning automatic metrics straightforward, as correlation with normalized adequacy scores can be used directly as an objective function for tuning.

3.2 Ranking Judgments

Introduced into the WMT evaluation in 2007 (Callison-Burch et al., 2007), the ranking task aims to remedy issues with adequacy and fluency by replacing arbitrary numeric scales with relative judgments. Given a reference translation and multiple translation hypotheses, annotators are asked to rank the translations from worst to best (allowing ties), a task facilitated by the availability of MT system outputs from the accompanying shared translation task. Translation ranking has the advantageous property that fine-grained distinctions can be made between translations that would not be possible in the adequacy task: sentences differing by single words or

phrases that would be forced into the same adequacy category can be easily ranked. This is especially important in evaluations where many similar MT systems compete, producing output that is nearly identical for many source sentences.

As with adequacy and fluency, annotator agreement in the ranking task can be evaluated with the kappa coefficient. Shown in Table 1, both inter-annotator and intra-annotator agreement are higher in the ranking task than for adequacy or fluency. Based on the results of WMT07, the ranking task is made the default form of human judgment in WMT08 (Callison-Burch et al., 2008) and WMT09 (Callison-Burch et al., 2009). Although Table 2 shows a general slight decline in annotator agreement over these evaluations, attributable to the increasing number of similar MT systems providing translation hypotheses, the kappa values remain relatively high.

Despite reported advantages, annotators still disagree on rankings in many cases and participants in WMT evaluations report several instances where ranking translations presents particular difficulty. Notably, the problem of longer sentences is even greater when annotators must keep multiple sentences in mind, leading to annotators' breaking down the task into a series of phrase-level judgments. As different judges cope with long sentences differently, many contradictory judgments are collected. Further, judges cannot reliably reproduce their own decompositions of longer sentences, leading to decreased intra-annotator agreement.

Even when annotators agree, particularly troublesome cases appear in evaluations where tens of similar systems compete, undermining ranking task informativeness. Consider three translations of a short sentence that are identical except for handling of some source word, for which they contain the following:

1. Source word translated incorrectly
2. Source word dropped
3. Foreign source word passed through

What is the correct ranking for these translations? Different errors are clearly present, but judges are largely unable to produce a consistent ordering and determining all three to be "ties" is uninformative. Similar problems frequently occur in larger sentences: difficult-to-compare errors exist in different numbers in each translation. Judges must decide which errors have the greatest impact on translation quality, an issue present in the adequacy task hoped to be avoided in ranking. Consider the following erroneous translations of some source sentence composed of phrases $p_1 ... p_4$:

1. p_1 translated incorrectly
2. p_2 translated incorrectly, where p_2 is half the length of p_1
3. p_3 and p_4 translated incorrectly, where combined length of p_3 and p_4 is less than length of p_1 or p_2
4. All content words correct but several function words missing
5. Main verb incorrectly negated

Again, many classes of errors are present, but producing a consistent ranking is incredibly difficult. While "tie" judgments can be used to determine that many systems are roughly equivalent, they remain unhelpful for error analysis of individual systems, a key use of human judgments. Further, "tie" judgments resulting from annotators' unanimous difficulty with certain sets of sentences can inflate annotator agreement, masking underlying problems with task informativeness.

Another difficulty arises when combining judgments from multiple annotators: while conflicting judgments in the adequacy task can be *averaged*, ideally to approximate "true" adequacy scores, conflicting ranking judgments actually *invalidate* one other. This is especially well illustrated in the case of tuning automatic metrics, a task which requires a clear objective function. The objective function for rankings presented in the WMT evaluations is *rank consistency*, the proportion of pairwise rankings preserved when translations are reranked according to metric scores. When two conflicting judgments exist, it is guaranteed that the metric will correctly replicate one and fail to replicate the other, invalidating both data points. Conflicting judgments could

be normalized by converting each pair of conflicting judgments into a single "tie" judgment, however this leads to the second problem of tuning to ranking judgments: as most metrics have no notion of a tie condition, tie judgments must be *discarded* prior to calculating rank consistency. Thus a large portion of collected data is unusable for this task and the addition of annotators actually *increases* the chance of data points being unusable. This is an inverse scenario of absolute rating tasks, in which multiple annotators *decrease* possibility of inaccurate data.

3.3 Post-Editing Judgments

Rather than directly eliciting absolute or relative judgments of translation quality, post-editing tasks attempt to measure the minimum amount of editing required by a human annotator to "fix" machine translation output. The most widely used post-editing measure is human-targeted translation edit rate (HTER) (Snover et al., 2006), in which annotators create *targeted* references by editing translation hypotheses to be fully meaning-equivalent with regular, non-targeted reference translations. The TER metric is then used to automatically calculate the number of edits between the original hypothesis and the targeted reference (edited hypothesis). To ensure HTER scores are as close as possible to actual minimum edit distance, annotators are instructed to use as few edits as possible when correcting translations. HTER is used as a primary measure of translation quality in the Global Autonomous Language Exploitation (GALE) program (Olive, 2005).

HTER addresses several problems arising in adequacy and ranking tasks. Foremost, as annotators do not assign any sort of rating, difficult decisions about what attributes are important for good translations or how harshly certain errors should be penalized are avoided entirely: post-editors must only correct translations to be semantically equivalent to references. Long sentences also pose less of a problem as they can be corrected incrementally rather than require a single blanket judgment. Finally, the post-editing process creates two useful byproducts: an additional set of reference translations and a set of edits pinpointing specific areas of incorrect translation. Both are highly useful for MT system development and error analysis.

The drawbacks of post-editing measures center on their reliance on automatic metrics to calculate edit distance. For example, HTER inherits the weaknesses of the TER measure (Snover et al., 2006): all deletions, insertions, and substitutions are treated equally. Incorrect forms of correct base words count as entire substitutions, no distinctions are made between content and function words, and negation is often reduced to a single insertion or deletion of a negation term. The corresponding advantage of using automatic measures centers on the fact that human edits must only be conducted once: newer, improved measures of translation distance can be rapidly applied to existing edit data sets.

As with the adequacy task, the availability of multiple annotators improves judgment accuracy. Rather than *averaging* the scores of multiple annotators, HTER takes the *minimum* score over all annotators as it is, by definition, the *minimum* edit distance. Tuning automatic evaluation metrics to HTER is also straightforward as numerical sentence-level scores are produced.

An additional post-editing task is introduced in WMT (Callison-Burch et al., 2009), in which post-editors are given translation hypotheses with *no* reference translations and asked to correct the translations to be fully *fluent*. A second task asks annotators whether or not edited hypotheses are meaning-equivalent with given reference translations. This two-stage task investigates the feasibility of using monolingual post-editors to correct MT output. To our knowledge, no work has yet utilized data from this task to develop automatic metrics.

4 Automatic Evaluation Metrics

Originally developed to stand in for human judgments in cases where collecting such judgments would be prohibitively time-consuming or expensive, automatic metrics of translation quality have many attractive properties. Not only do metrics score data sets quickly, but the problems of annotator agreement are not encountered as most metric scoring algorithms are deterministic. During minimum error rate training (MERT) (Och, 2003), many nearly-identical hypotheses must be reliably scored in a short amount of time. During error analysis, a single feature might be added or subtracted from a translation system, resulting in changes barely de-

tectable by humans. In such cases, *any* annotator disagreement can undermine the informativeness of judgment data.

To be effective, metrics must also have high correlation with the human judgments they are standing in for. To accomplish this, many recent metrics include several parameters that can be tuned to maximize correlation with various types of judgments. This leads to the questions of whether or not metrics can be tuned reliably and what sorts of judgments are ideal tuning data. Section 5 discusses experiments of metric tuning for several judgment types using METEOR-NEXT, a highly tunable metric.

4.1 METEOR-NEXT

The METEOR-NEXT metric (Denkowski and Lavie, 2010) evaluates a machine translation hypothesis against a reference translation by calculating a score based on a phrase alignment between the two sentences. If multiple reference translations are available, the hypothesis is scored against each and the reference producing the highest final score is used.

For each hypothesis-reference pair, an alignment is constructed between the two sentences in a two stage process. In *stage one*, all possible word and phrase matches between the sentences are identified according to the following matchers:

Exact: Words are matched if and only if their surface forms are identical.

Stem: Words are stemmed using a Snowball Stemmer (Porter, 2001) and matched if the stems are identical.

Synonym: Words are matched if they share membership in a synonym set according to the WordNet (Miller and Fellbaum, 2007) database.

Paraphrase: Phrases are matched if they are listed as paraphrases in the METEOR paraphrase tables. These paraphrase tables are constructed by applying the techniques described by Callison-Burch (2005) to portions of the shared translation task data available for the 2010 ACL Workshop on Statistical Machine Translation (Callison-Burch et al., 2010).

Although the exact, stem, and synonym matchers identify *word* matches while the paraphrase matcher identifies *phrase* matches, all matches are generalized to phrase matches with both a start position and phrase length in each sentence. A word occurring less than *length* positions after a match start is considered *covered* by the match. Exact, stem, and synonym matches always cover one word in each sentence while paraphrase matches can cover one or more words in either sentence.

In *stage two*, the final alignment is identified as the largest subset of all matches meeting the following criteria in order of importance:

1. Require each word in each sentence to be covered by *at most* one match.
2. Choose the largest number of words covered across *both* sentences.
3. Choose the smallest number of chunks, where a *chunk* is defined as a contiguous series of matched phrases that is identically ordered in both sentences.
4. Choose the smallest sum of absolute distances between match start positions in the two sentences. (Break ties by preferring to align words and phrases that occur at similar positions in both sentences.)

Once an alignment is constructed, the METEOR-NEXT score is calculated as follows. The number of words in the translation hypothesis (t) and reference translation (r) are counted. For each of the matchers (m_i), count the number of words covered by matches of this type in the hypothesis ($m_i(t)$) and reference ($m_i(r)$). Use the matcher weights (w_i) to calculate the weighted Precision and Recall:

$$P = \frac{\sum_i w_i \cdot m_i(t)}{|t|} \qquad R = \frac{\sum_i w_i \cdot m_i(r)}{|r|}$$

The parameterized harmonic mean of P and R (van Rijsbergen, 1979) is then calculated:

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1 - \alpha) \cdot R}$$

To account for gaps in translation and differences in word order, a fragmentation penalty is calculated using the total number of matched words (m) and number of chunks (ch):

$$Pen = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}$$

The final METEOR-NEXT score is then calculated:

$$Score = (1 - Pen) \cdot F_{mean}$$

The parameters α, β, γ, w_{exact}, w_{stem}, $w_{synonym}$, and $w_{paraphrase}$ can be tuned to maximize correlation with human judgments.

5 Experiments

To explore both human evaluation tasks and the task of tuning automatic evaluation metrics, we tune versions of METEOR-NEXT on various human judgment data sets. Following Snover et al. (2009), we examine the optimal parameter values for each type of human judgment. We also examine the correlation of each METEOR-NEXT version with human judgments from all other sets to determine the relative benefit of tuning to various types of human judgments. Correlation results for METEOR-NEXT are compared to those for three baseline metrics: BLEU (Papineni et al., 2002), TER (Snover et al., 2006), and METEOR (Lavie and Agarwal, 2007).

5.1 Data

We conduct tuning and evaluation experiments on six data sets spanning three human judgment tasks over two consecutive years. **Adequacy** data includes the full NIST Open MT human evaluation sets for (1) 2008 (Przybocki, 2008) and (2) 2009 (Przybocki, 2009). **Ranking** data includes all WMT ranking judgments for translations into English for (1) 2008 (Callison-Burch et al., 2008) and (2) 2009 (Callison-Burch et al., 2009). **HTER** data includes the GALE (Olive, 2005) unsequestered human evaluation data for (1) Phase 2 and (2) Phase 3. Where possible, we use the same data and evaluation criteria as major evaluations so that reported scores are comparable.

5.2 Tuning Procedure

For the tasks of adequacy and HTER, METEOR-NEXT parameters are tuned to maximize the sentence-level length-weighted Pearson's correlation of METEOR-NEXT scores with human judgments. Following Callison-Burch et al. (Callison-Burch et al., 2009), ranking versions of METEOR-NEXT are tuned to maximize rank consistency, the proportion of pairwise ranking judgments preserved when hypotheses are reranked by metric score. All "tie" judgments are discarded prior to tuning. In all cases, parameters are tuned via exhaustive grid search of feasible parameter space. The resulting *globally* optimal parameters are ideal for examining the characteristics of each judgment type.

5.3 Results

Table 3 shows the optimal parameters for each of the six tuning sets. Notably, parameters for the adequacy and ranking tasks fluctuate between years while parameters for HTER are most stable. Especially ranking evaluations containing many nearly-identical translation hypotheses can dramatically skew parameters values. To achieve maximum rank consistency, parameters are chosen to severely penalize small differences while remaining indifferent to larger positive or negative qualities shared by all translations. For example, if translations differ only by word form, stem matches might receive a zero weight, or if translations differ only by word order, the fragmentation penalty might receive majority weight. These parameters are unhelpful for ranking other sets of hypotheses with different slight differences between them. Similar issues can occur in adequacy tasks when translations hypotheses from two different years have different, difficult-to-compare errors.

While all parameter sets favor recall over precision, the ranking task is a particularly extreme case, followed by the adequacy task. In the case of adequacy, this can be attributed to annotators' tendency to first read reference translations and look for the same information in hypotheses. In the case of ranking, most MT systems are tuned using the precision-based BLEU metric and are thus more likely to differ in recall. Post-editing forces annotators to consider *both* missing and extraneous information by editing hypotheses to have the exact meaning of references, leading to more balance parameters for HTER. Also notable is the HTER task's low weight for stem matches, caused by the TER metric's lack of such matches. Finally, the ranking task has the slightest fragmentation penalty, reflecting the highly similar word order of ranked hypotheses, followed by the adequacy task, while the HTER task has the harshest penalty, reflecting the strict requirement that each reordering requires an edit to correct.

Table 4 shows the correlation and rank consistency results for METEOR-NEXT versions tuned on each type of data, as well as results for baseline metrics BLEU (Papineni et al., 2002), TER (Snover et

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

Table 3: Optimal METEOR-NEXT parameter values for several human judgment data sets.

		Adequacy (r)		Ranking (consist)		HTER (r)	
Metric	Tuning Data	MT08	MT09	WMT08	WMT09	GALE-P2	GALE-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
METEOR-NEXT	MT08	*0.620*	0.625	0.630	0.614	**-0.638**	-0.590
METEOR-NEXT	MT09	0.612	*0.630*	**0.637**	0.617	-0.636	-0.589
METEOR-NEXT	WMT08	0.598	**0.626**	*0.643*	**0.621**	-0.629	-0.573
METEOR-NEXT	WMT09	0.601	0.624	0.635	*0.629*	-0.628	-0.578
METEOR-NEXT	GALE-P2	**0.616**	0.623	0.632	0.615	*-0.640*	**-0.596**
METEOR-NEXT	GALE-P3	0.610	0.618	0.636	0.617	**-0.638**	*-0.600*

Table 4: Sentence-level Pearson's r and rank consistency of metrics with human judgments on MT evaluation data sets. Italics indicate METEOR-NEXT tuned on given data set (oracle performance) and bold indicates highest scoring metric tuned on other data. Dashes indicate no sentence-level data available for metric on given data set.

al., 2006), and METEOR (Lavie and Agarwal, 2007). Notably, all versions of METEOR-NEXT outperform all three baseline metrics on every data set, indicating that METEOR-NEXT is a highly stable metric capable of achieving similar correlation levels with various parameter sets. Also notable is the effectiveness of tuning to HTER; in addition to being the only task for which tuning on an alternate year's data consistently produces the highest correlation with judgments, HTER parameters also achieve similar or higher correlation levels on other types of judgments than parameter sets tuned on the same type of judgment from alternate years. The adequacy task also does well in this regard, though slightly less so than HTER, while the performance of ranking parameters is inconsistent. This generally follows the trend of parameter balance and stability in Table 3.

6 Conclusions

We have examined several types of human judgment tasks across criteria such as performance of annotators, informativeness of results, and practicality of use in automatic metric development. Adequacy tasks present several sources of confusion and difficulty for annotators, although the addition of multiple annotators can mitigate these issues. While ranking judgments address some problems with the adequacy task, unintended complications arise when translations are either very long or contain multiple difficult-to-compare errors and the addition of multiple annotators can actually invalidate data. Post-editing tasks shift scoring responsibility entirely to automatic metrics; while many causes of difficulty encountered in adequacy and fluency tasks are avoided, high quality automatic measures are required to ensure score accuracy.

Our tuning experiment reveals that the most stable metric parameters are achieved when tuning to HTER data, and that HTER-tuned parameters produce the best overall correlation results, followed closely by adequacy-tuned parameters. Further, the fact that all METEOR-NEXT correlation scores fall within a small range well above the scores of baseline metrics indicates that METEOR-NEXT is a sta-

ble metric and practical choice for tuning to various types of human judgments.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proc. of ACL05*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence Estimation for Machine Translation. Technical Report Natural Language Engineering Workshop Final Report, Johns Hopkins University.

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.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *Proc. Third Workshop on Statistical Machine Translation*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proc. Fourth Workshop on Statistical Machine Translation*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan. 2010. Fifth Workshop on Statistical Machine Translation. http://www.statmt.org/wmt10/.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR Paraphrase Tables: Improve Evaluation Support for Five Target Languages. In *Proc. ACL WMT/MetricsMATR10*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proc. of WMT07*, pages 228–231.

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

George Miller and Christiane Fellbaum. 2007. WordNet. http://wordnet.princeton.edu/.

Franz Josef Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Martin Porter. 2001. Snowball: A language for stemming algorithms. http://snowball.tartarus.org/texts/.

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/.

Mark Przybocki. 2009. NIST Open Machine Translation 2009 Evaluation. http://www.itl.nist.gov/iad/mig/tests/mt/2009/.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. 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)*, pages 223–231.

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*.

C. van Rijsbergen, 1979. *Information Retrieval*, chapter 7. 2nd edition.