Exploring Normalization Techniques for Human Judgments of Machine Translation Adequacy Collected Using Amazon Mechanical Turk

Michael Denkowski and **Alon Lavie**
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15232, USA
{mdenkows,alavie}@cs.cmu.edu

Abstract

This paper discusses a machine translation evaluation task conducted using Amazon Mechanical Turk. We present a translation adequacy assessment task for untrained Arabic-speaking annotators and discuss several techniques for normalizing the resulting data. We present a novel 2-stage normalization technique shown to have the best performance on this task and further discuss the results of all techniques and the usability of the resulting adequacy scores.

1 Introduction

Human judgments of translation quality play a vital role in the development of effective machine translation (MT) systems. Such judgments can be used to measure system quality in evaluations (Callison-Burch et al., 2009) and to tune automatic metrics such as METEOR (Banerjee and Lavie, 2005) which act as stand-ins for human evaluators. However, collecting reliable human judgments often requires significant time commitments from expert annotators, leading to a general scarcity of judgments and a significant time lag when seeking judgments for new tasks or languages.

Amazon's Mechanical Turk (MTurk) service facilitates inexpensive collection of large amounts of data from users around the world. However, Turkers are not trained to provide reliable annotations for natural language processing (NLP) tasks, and some Turkers attempt to game the system by submitting random answers. For these reasons, NLP tasks must be designed to be accessible to untrained users and data normalization techniques must be employed to ensure that the data collected is usable.

This paper describes a MT evaluation task for translations of English into Arabic conducted using MTurk and compares several data normalization techniques. A novel 2-stage normalization technique is demonstrated to produce the highest agreement between Turkers and experts while retaining enough judgments to provide a robust tuning set for automatic evaluation metrics.

2 Data Set

Our data set consists of human adequacy judgments for automatic translations of 1314 English sentences into Arabic. The English source sentences and Arabic reference translations are taken from the Arabic-English sections of the NIST Open Machine Translation Evaluation (Garofolo, 2001) data sets for 2002 through 2005. Selected sentences are between 10 and 20 words in length on the Arabic side. Arabic machine translation (MT) hypotheses are obtained by passing the English sentences through Google's free online translation service.

2.1 Data Collection

Human judgments of translation adequacy are collected for each of the 1314 Arabic MT output hypotheses. Given a translation hypothesis and the corresponding reference translation, annotators are asked to assign an adequacy score according to the following scale:

4 – Hypothesis is completely meaning equivalent with the reference translation.

Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 57–61,
Los Angeles, California, June 2010.

3 – Hypothesis captures more than half of meaning of the reference translation.

2 – Hypothesis captures less than half of meaning of the reference translation.

1 – Hypothesis captures no meaning of the reference translation.

Adequacy judgments are collected from untrained Arabic-speaking annotators using Amazon's Mechanical Turk (MTurk) service. We create a human intelligence task (HIT) type that presents Turkers with a MT hypothesis/reference pair and asks for an adequacy judgment. To make this task accessible to non-experts, the traditional definitions of adequacy scores are replaced with the following: (4) excellent, (3) good, (2) bad, (1) very bad. Each rating is accompanied by an example from the data set which fits the corresponding criteria from the traditional scale. To make this task accessible to the Arabic speakers we would like to complete the HITs, the instructions are provided in Arabic as well as English.

To allow experimentation with various data normalization techniques, we collect judgments from 10 unique Turkers for each of the translations. We also ask an expert to provide "gold standard" judgments for 101 translations drawn uniformly from the data. These 101 translations are recombined with the data and repeated such that every 6th translation has a gold standard judgment, resulting in a total of 1455 HITs. We pay Turkers \$0.01 per HIT and Amazon fees of \$0.005 per HIT, leading to a total cost of \$218.25 for data collection and an effective cost of \$0.015 per judgment. Despite requiring Arabic speakers, our HITs are completed at a rate of 1000-3000 per day. It should be noted that the vast majority of Turkers working on our HITs are located in India, with fewer in Arabic-speaking countries such as Egypt and Syria.

3 Normalization Techniques

We apply multiple normalization techniques to the data set and evaluate their relative performance. Several techniques use the following measures:

- Δ: For judgments ($J = j_1...j_n$) and gold standard ($G = g_1...g_n$), we define average distance:

$$\Delta(J, G) = \frac{\sum_{i=1}^{n} |g_i - j_i|}{n}$$

- K: For two annotators, Cohen's kappa coefficient (Smeeton, 1985) is defined:

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

where $P(A)$ is the proportion of times that annotators agree and $P(E)$ is the proportion of times that agreement is expected by chance.

3.1 Straight Average

The baseline approach consists of keeping all judgments and taking the straight average on a per-translation basis without additional normalization.

3.2 Removing Low-Agreement Judges

Following Callison-Burch et al. (2009), we calculate pairwise inter-annotator agreement ($P(A)$) of each annotator with all others and remove judgments from annotators with $P(A)$ below some threshold. We set this threshold such that the highest overall agreement can be achieved while retaining at least one judgment for each translation.

3.3 Removing Outlying Judgments

For a given translation and human judgments ($j_1...j_n$), we calculate the distance (δ) of each judgment from the mean ($\bar{j}$):

$$\delta(j_i) = |j_i - \bar{j}|$$

We then remove outlying judgments with $\delta(j_i)$ exceeding some threshold. This threshold is also set such that the highest agreement is achieved while retaining at least one judgment per translation.

3.4 Weighted Voting

Following Callison-Burch (2009), we treat evaluation as a weighted voting problem where each annotator's contribution is weighted by agreement with either a gold standard or with other annotators. For this evaluation, we weigh contribution by $P(A)$ with the 101 gold standard judgments.

3.5 Scaling Judgments

To account for the notion that some annotators judge translations more harshly than others, we apply per-annotator scaling to the adequacy judgments based on annotators' signed distance from gold standard judgments. For judgments ($J = j_1...j_n$) and gold standard ($G = g_1...g_n$), an additive scaling factor is calculated:

$$\lambda_+(J,G) = \frac{\sum_{i=1}^{n} g_i - j_i}{n}$$

Adding this scaling factor to each judgment has the effect of shifting the judgments' center of mass to match that of the gold standard.

3.6 2-Stage Technique

We combine judgment scaling with weighted voting to produce a 2-stage normalization technique addressing two types of divergence in Turker judgments from the gold standard. Divergence can be either consistent, where Turkers regularly assign higher or lower scores than experts, or random, where Turkers guess blindly or do not understand the task.

Stage 1: Given a gold standard ($G = g_1...g_n$), consistent divergences are corrected by calculating $\lambda_+(J,G)$ for each annotator's judgments ($J = j_i...j_n$) and applying $\lambda_+(J,G)$ to each j_i to produce adjusted judgment set J'. If $\Delta(J',G) < \Delta(J,G)$, where $\Delta(J,G)$ is defined in Section 3, the annotator is considered consistently divergent and J' is used in place of J. Inconsistently divergent annotators' judgments are unaffected by this stage.

Stage 2: All annotators are considered in a weighted voting scenario. In this case, annotator contribution is determined by a distance measure similar to the kappa coefficient. For judgments ($J = j_1...j_n$) and gold standard ($G = g_1...g_n$), we define:

$$K_\Delta(J,G) = \frac{(\max\Delta - \Delta(J,G)) - E(\Delta)}{\max\Delta - E(\Delta)}$$

where $\max\Delta$ is the average maximum distance between judgments and $E(\Delta)$ is the expected distance between judgments. Perfect agreement with the gold standard produces $K_\Delta = 1$ while chance agreement produces $K_\Delta = 0$. Annotators with $K_\Delta \leq 0$ are removed from the voting pool and final scores are calculated as the weighted averages of judgments from all remaining annotators.

Type	Δ	K_Δ
Uniform-a	1.02	0.184
Uniform-b	1.317	-0.053
Gaussian-2	1.069	0.145
Gaussian-2.5	0.96	0.232
Gaussian-3	1.228	0.018

Table 2: Weights assigned to random data

4 Results

Table 1 outlines the performance of all normalization techniques. To calculate $P(A)$ and K with the gold standard, final adequacy scores are rounded to the nearest whole number. As shown in the table, removing low-agreement annotators or outlying judgments greatly improves Turker agreement and, in the case of removing judgments, decreases distance from the gold standard. However, these approaches remove a large portion of the judgments, leaving a skewed data set. When removing judgments, 1172 of the 1314 translations receive a score of 3, making tasks such as tuning automatic metrics infeasible.

Weighing votes by agreement with the gold standard retains most judgments, though neither Turker agreement nor agreement with the gold standard improves. The scaling approach retains all judgments and slightly improves correlation and Δ, though K decreases. As scaled judgments are not whole numbers, Turker $P(A)$ and K are not applicable.

The 2-stage approach outperforms all other techniques when compared against the gold standard, being the only technique to significantly raise correlation. Over 90% of the judgments are used, as shown in Figure 1. Further, the distribution of final adequacy scores (shown in Figure 2) resembles a normal distribution, allowing this data to be used for tuning automatic evaluation metrics.

4.1 Resistance to Randomness

To verify that our 2-stage technique handles problematic data properly, we simulate user data from 5 unreliable Turkers. Turkers "Uniform-a" and "Uniform-b" draw answers randomly from a uniform distribution. "Gaussian" Turkers draw answers randomly from Gaussian distributions with $\sigma = 1$ and μ according to name. Each "Turker" contributes one judgment for each translation. As shown in Ta-

Technique	Retained	Gold Standard				Turker	
		Correlation	Δ	$P(A)$	K	$P(A)$	K
Straight Average	**14550**	0.078	0.988	0.356	0.142	0.484	0.312
Remove Judges	6627	-0.152	1.002	0.347	0.129	0.664	0.552
Remove Judgments	9250	0	0.891	0.356	0.142	**0.944**	**0.925**
Weighted Voting	14021	0.152	0.968	0.356	0.142	0.484	0.312
Scale Judgments	**14550**	0.24	0.89	0.317	0.089	N/A	N/A
2-Stage Technique	13621	**0.487**	**0.836**	**0.366**	**0.155**	N/A	N/A

Table 1: Performance of normalization techniques

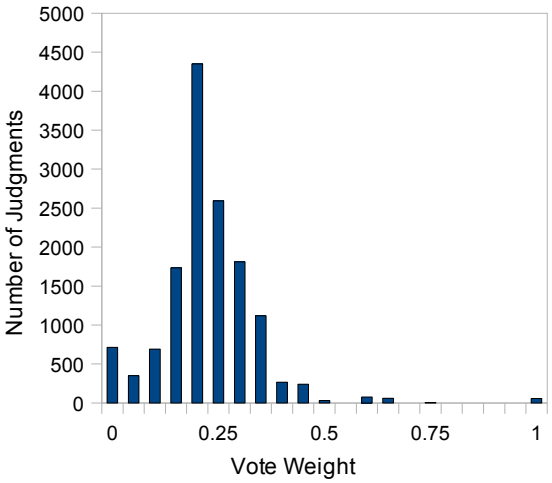

Figure 1: Distribution of weights for judgments

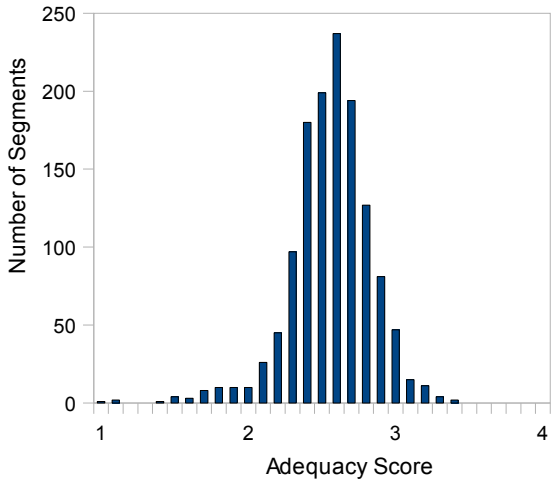

Figure 2: Distribution of adequacy scores after 2-stage normalization

ble 2, only Gaussian-2.5 receives substantial weight while the others receive low or zero weight. This follows from the fact that the actual data follows a similar distribution, and thus the random Turkers have negligible impact on the final distribution of scores.

5 Conclusions and Future Work

We have presented an Arabic MT evaluation task conducted using Amazon MTurk and discussed several possibilities for normalizing the collected data. Our 2-stage normalization technique has been shown to provide the highest agreement between Turkers and experts while retaining enough judgments to avoid problems of data sparsity and appropriately down-weighting random data. As we currently have a single set of expert judgments, our future work involves collecting additional judgments from multiple experts against which to further test our techniques. We then plan to use normalized Turker adequacy judgments to tune an Arabic version of the METEOR (Banerjee and Lavie, 2005) MT evaluation metric.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proc. ACL WIEEMMTS*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of WMT09. In *Proc. WMT09*.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proc. EMNLP09*.

John Garofolo. 2001. NIST Open Machine Translation Evaluation. http://www.itl.nist.gov/iad/mig/tests/mt/.

N. C. Smeeton. 1985. Early History of the Kappa Statistic. In *Biometrics*, volume 41.

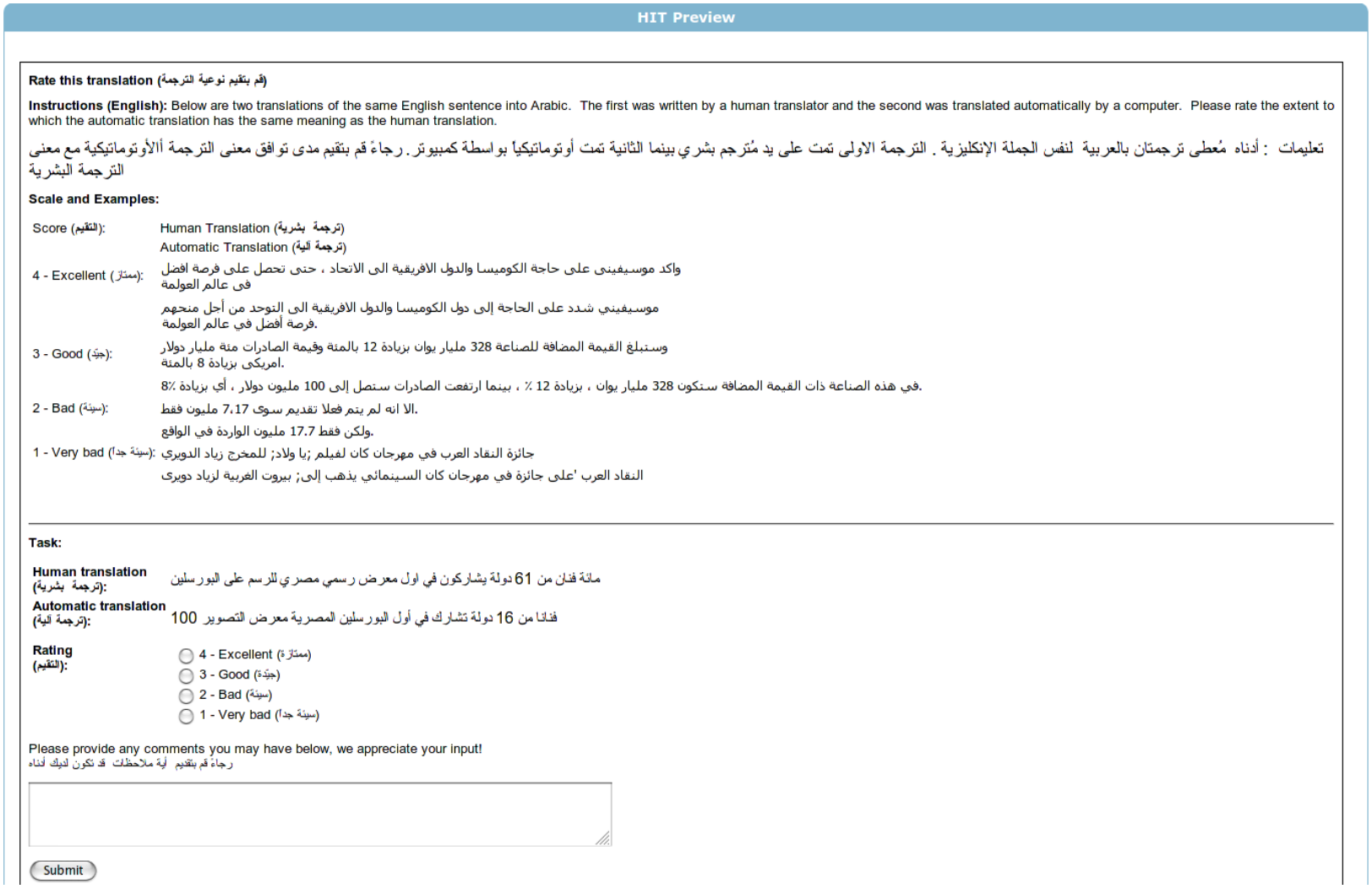

Figure 3: Example HIT as seen by Turkers