# Evaluating Human Correction Quality for Machine Translation from Crowdsourcing

**Shasha Liao**
Computer Science Department
New York University
liaoss@cs.nyu.edu

**Cheng Wu, Juan Huerta**
IBM T.J. Watson Research Center
{chengwu,huerta}@us.ibm.com

## Abstract

Machine translation (MT) technology is becoming more and more pervasive, yet the quality of MT output is still not ideal. Thus, human corrections are used to edit the output for further studies. However, how to judge the human correction might be tricky when the annotators are not experts. We present a novel way that uses cross-validation to automatically judge the human corrections where each MT output is corrected by more than one annotator. Cross-validation among corrections for the same machine translation, and among corrections from the same annotator are both applied. We get a correlation around 40% in sentence quality for Chinese-English and Spanish-English. We also evaluate the user quality as well. At last, we rank the quality of human corrections from good to bad, which enables us to set a quality threshold to make a trade-off between the scope and the quality of the corrections.

## 1 Introduction

Human corrections are aimed to give the correct translation by editing the MT output. In this way, they can be used to analyze what kind of mistakes a MT system might make; also, they can be feed back to the MT system to improve the output. Manual human correction is generally thought to be excessively time consuming and expensive and experts are acquired to make sure the quality. However, as the scale of online multi-user communities is increasing, it becomes an easier and faster way to collect a large amount of human corrections. Crowdsourcing is an effective and cost-efficient way to collect human corrected (HC) sentences. However, before feeding back the crowdsourcing data to MT system, there are two challenges (a) how to measure the quality of HC sentences and (b) how to select good quality HC sentences for enhancing the translation models.

If we only have one human correction per sentence, the quality is quite hard to evaluate. However, if each sentence contains more than one human correction, there are much more information we are able to use. In this paper, we used the redundant corrections to apply a cross-validation approach to automatically evaluate the human corrections and rank them.

## 2 Crowdsourcing Description

Our crowdsourcing is based on enterprise data from employee participation in translation tasks, and conducted inside a worldwide company (Osamuyimen Stewart etc. 2010). This is used to help us with the data collection effort required for improving statistical machine translation algorithms, by harnessing the linguistic skills of worldwide bi-lingual employees for accomplishing the complex translation task that is typically done by professional translators. Participants are presented with text of relevant data e.g., news, technical content, history, etc., in a source language and asked to translate into a target language.

In this paper, we use "benchmark" data in crowdsourcing, where each source sentence contains multiple human corrections from multiple participants. In Benchmark data, for each source sentence, we collect one machine translation sentence, more than two human corrections. Each set is called a *translation set,* and experts are asked to give one reference for each *translation set*[1].

## 3 Cross-validation for Sentences

In this section, we proposed the cross-validation method. Different features will be examined and

[1] Note that our goal is to evaluate human corrections without reference, and the reference is not used in our



*Proceedings of Recent Advances in Natural Language Processing*, pages 598–603,
Hissar, Bulgaria, 12-14 September 2011.

combined to reach the best performance. The basic assumption of cross-validation is that if a correction is similar to other corrections in the same translation set, it implies that other annotators probably agree with this correction; otherwise, it means other annotators has very different opinion on this correction. Thus, if a correction has high similarities to other corrections, it is probably a good correction; otherwise, it is not. By applying this "pseudo-reference" approach, we can judge the sentence level quality more confidently.

As there are many different methods to calculate the similarity, on lexical, syntax, or even semantic levels, we apply these features and evaluate them in the rest of this section. We first apply BLEU, the traditional metric for MT evaluation, and then other features including word similarity, semantic analysis and syntax information, which are widely used in NLP tasks.

Also, based on the special characteristics of the crowdsourcing, we also apply another similarity from the "user" view, where the user quality is used as a special feature.

### 3.1 Language Model (LM)

Language model are used a lot in machine translation. The basic assumption is that a good translation should be more fluent, and more like the standard sentences.

SRI language model toolkit[2] is used to train the language models from 68,101 English sentences in Crowdsourcing which are translated to other languages, and a 5-gram language models is built. The perplexity score is normalized by the largest perplexity score in the *translation set*.

$$LM^*(s) = \frac{ppl(\max(Set(s))}{ppl(s)}$$

Where *Set(s)* is the translation set containing s, and *max(Set(s))* is the maximal language score in *Set(s), ppl(s)* is the perplexity score for sentence s[3].

### 3.2 Cross-validated Bleu Score (c-Bleu)

First, we apply a straightforward strategy to see if we can only use BLEU score among different human corrections from the same translation set to give cross-validated score. In this method, only the n-gram among sentences is used, and no linguistic knowledge is needed. Thus, this is a very convenient method, and can apply to evaluation on translation from any language pairs.

### 3.3 Cross-validated Word Similarity (c-WS)

Instead of using BLEU score, we apply another method of evaluating the translation by calculating the similarity between two sentences. Tokenization is applied before calculation, and the word order is not considered in the normal word similarity. Every sentence is treated as a word vector $Si = (w_{i1}, w_{i2}, ..., w_{in})$, and for two sentences $S_1$ and $S_2$, the similarity between them is:

$$word\_sim(S_1, S_2) = \frac{\sum_{w_{1i} \in S1, w_{2j} \in S2} sim(w_{1i}, w_{2j})}{\sqrt{|S_1| * |S_2|}}$$

Where $sim(w_{1i}, w_{2j})$ equals 1 if $w_{1i}$ equals $w_{2j}$, otherwise 0.

#### 3.3.1 Cross-validated Stemmed WS (c-WS1)

Some languages like Chinese don't have plural for example, and translator might translate a Chinese word with single or plural form, which are both correct. This also happens for past form and present form too. As a result, we test another similarity metric that ignores such difference. For example, "attacked" will be stemmed to "attack", and "rules" will be stemmed to "rule". However, as different word forms might predicate different functions in the sentence, for two different words with the same base form, we give them a similarity score of 0.95.

#### 3.3.2 Cross-validated Semantic WS (c-WS2)

Translation can be very different, and people might use different word with the same meaning. For example, "search" and "find" are both good translation for Chinese word "查找". This is also one important reason why evaluation with multiple references is better than that with single reference. In this metric, we involve such information to calculate the similarity between two corrections. In practice, we use WordNet[4] to calculate the semantic similarity between two words (Leacock and Chodorow 1998, Wu and Palmer 1994, Resnik 1995, Lin 1998, and Jiang and Conrath 1997). In our experiment, we use the Information Content (IC) method presented

[2] http://www.speech.sri.com/projects/srilm/

[3] Note that a better sentence will have a lower perplexity score, and we use the inverse of ppl as the language model score.

[4] http://wordnet.princeton.edu/

by Lin (1998), where the IC score ranges from 0.0 to 1.0.

For a sentence, the semantic word similarity between S1 and S2 is calculated by:

```
senscore = 0.0;
For each word w1 in S1
   best_match = 0
   For each word w2 in S2
      score = SimFun(w1,w2)
      If score > best_match
         best_match = score
   if(best_match > threshold)
      senscore += best_match
      remove w1 from S1, remove w2 from S2
senscore = senscore/sqrt(length(S1)*length(S2))
```

Figure1. Procedure of computing semantic similarity for two sentences

#### 3.3.3 Cross-validated Syntax-based WS (c-WS3)

In the above similarity method, the relations between words are not considered, thus no syntax information is provided. In this similarity method, we want to take the dependency tree similarity into consideration. In our experiment, we use the Stanford Dependencies[5] to acquire such syntactic information.

We use the triplets in the dependency tree, which is composed of (relation, governor, dependent). And for every pair of triplets (t1, t2), we calculate its similarity in this way

$$dep\_sim(t1,t2) = sim(relation_{t1}, relation_{t2}) * sim(governor_{t1}, governor_{t2}) * sim(dependent_{t1}, dependent_{t2})$$

where relation will be 1 for exact match, and 0 otherwise. For governor and dependent, we use the semantic similarity mentioned in section 4.3.2. The similarity between two sentences is:

$$dep\_sim(S_1, S_2) = \frac{\sum_{t_{1i} \in T1, t_{2j} \in T2} sim(t_{1i}, t_{2j})}{\sqrt{|T_1| * |T_2|}}$$

However, syntax-based similarity is more sparse than word-based similarity, and we use a parameter $\alpha$ to balance between the two[6]:

$$Sim(S_1, S_2) = \alpha * word\_sim(S1, S2) + (1-\alpha) dep\_sim(S1, S2)$$

### 3.4 Cross-validated Correction Similarity (c-CS)

As the human correction is derived from machine translation, the difference between the correction and the translation might be more likely to reflect the quality of the corrections. As a result, we calculate the similarity between the corrections (adding and deleting) from machine translation instead of the whole sentence. The difference between sentence similarity and correction similarity is that: for sentence similarity, every sentence is represented by all the words in the sentence, while in correction similarity, we only consider about the words which are inserted or deleted from the machine translation. We also test the correction similarity on stemmed (c-CS1), semantic (c-CS2), and syntax level (c-CS3).

## 4 Cross-validation for User Evaluation

Above features treat each *translation set* as a whole, and user information are ignored. However, we believe that the user information can also be predictable. If a user's translation skill is good, he should always provide good corrections, while a user with limited translation skill will provide relatively worse corrections. Thus, if we can acquire user quality, we can use it to evaluate the sentence he corrects.

Although the user quality cannot be implicitly evaluated since we do not do any quality test, we can indirectly acquire such information based on the quality of the sentence he translates. As the user quality is judged by all the sentences he corrected, it should be more reliable even the evaluation on sentence is not very confident. The user score is calculated by:

$$US(u) = \frac{\sum_{si \in set(u)} score_{si}}{|set(u)|}$$

where $Set(u)$ is the set of sentence translated by user $u$,and $score_{si}$ is the sentence score calculated in section 3.3.

After we evaluate each user, we also feed it back as an extra feature for sentence level

[5] http://nlp.stanford.edu/software/stanford-dependencies.shtml

[6] In practice, we set $\alpha$ to 0.8.

evaluation. It is another kind of cross-validation, where the quality of a correction is based on other corrections from the same user.

## 5 Experiment

We use two methods for sentence level evaluation: one is a correlation evaluation that checks the correlation between different features and human assessment; the other is a selection evaluation to see if we can select good human corrections above a threshold to feedback to MT system.

We only use correlation evaluation for user quality evaluation, as we do not want to set a threshold to forbid any user to contribute.

We start with Chinese-English (C-E) and Spanish-English (S-E) MT. In C-E, there are 67 translation sets, with 335 human corrections and 39 people participated; while in S-E, there are 40 translation sets, with 217 human corrections and 38 people participated. Most translations are corrected 3 to 5 times. Users corrected different amount of sentences: some corrected one sentence, while some might correct more than 25 sentences.

### 5.1 Golden Standard

In this section, we create a key set by human assessment as our gold standard: we mixed up the machine translation, human corrections, and reference for one original sentence and 5 annotators in Chinese-English, and 3 in Spanish-English, were asked to assess the sentences scores from 1 to 5, where 1 corresponds to poor and 5 corresponds to perfect translation.

We calculate the average score of machine translation, reference, and human correction to see how good they are (table 1).

| | MT | Ref | HC* | H_HC |
|---|---|---|---|---|
| Chinese | 2.08 | 4.52 | *4.2* | *4.67* |
| Spanish | 2.96 | 4.3 | *3.9* | *4.5* |

Table1. Chinese-English overall qualities for machine translation, reference, and human correction[7]

### 5.2 Correlation Experiments

The basic assumption of correlation experiment is that a good evaluation metric should correlate better to the golden standard. We test the correlation of each feature to the human assessment, and also try to combine the features together to achieve the best performance. As no machine learning involved in this paper, we use simple multiplication to combine scores from different features.

Because we have a reference in benchmark data, we use the correlation between the bleu score between the correction and the reference as our baseline, which is not quite good.

#### 5.2.1 Sentence Evaluation Results

From table2 we can see that language model score correlates worst. This indicates that distinguishing human correction and machine learning might be easier, but distinguishing between corrections is much harder.

Cross validation on BLEU scores works better than bleu score with single reference, but it does not work as well as word similarity method.

Similarity calculation works best, and if more linguistic information is involved, the correlation is better. We try to combine different features together, and only report the ones that improve.

| Sentence Correlation / Methods | Chinese | Spanish |
|---|---|---|
| Baseline | 28.7% | 18.7% |
| c-Bleu | 29.7% | 24% |
| LM | 17.4% | -0.84% |
| c-WS | 33.7% | 31.7% |
| +Stemmed (c-WS1) | 35% | 32.8% |
| +Semantic (c-WS2) | 36.3% | 30.5% |
| +Syntax-based (c-WS3) | **37%** | **33.3%** |
| c-CS | 30.7% | 36.2% |
| +Stemmed (c-CS1) | 31.8% | **36.6%** |
| +Semantic (c-CS2) | 31.9% | 34.3% |
| +Syntax-based (c-CS3) | **33.1%** | 35.2% |
| c-WS3*c-CS3 | **38.8%** | **39.8%** |

Table2. Sentence correlation results for different features

#### 5.2.2 User Evaluation Results

The golden standard for each user is judged by the average quality of his corrections, and we test

[7] HC* is the overall HC score, and H_HC represents the best HC from each translation set

the correlation between golden standard and automatic evaluation (table 3).

| Methods \ User Correlation | Chinese | Spanish |
|---|---|---|
| c-WS3 | 52.2% | 51% |
| c-CS3 | 54.6% | 64.1% |
| c-WS3* c-CS3 | **57.8%** | **70.5%** |

Table3. Results of user quality correlation

Then we added user quality as an extra feature for sentence evaluation. Experiment shows that adding this feature can further improve the correlation by 1.8% for C-E and 1.6% for S-E (table 4).

| Methods \ Sentence Correlation | Chinese | Spanish |
|---|---|---|
| User_Score (US) | 30.6% | 28.5% |
| c-WS3* c-CS3 | 38.8% | 39.8% |
| c-WS3* c-CS3* US | **40.6%** | **41.4%** |

Table4. Results of feedback user quality to sentence quality

### 5.2.3 Analysis

From the study above, we can see that the similarity score among human corrections performs best, and it can achieve a better result than using bleu score with reference.

N-gram based language model does not help too much, but long distance features, like syntax feature, when combined with word similarity, is helpful.

Language model does not correlate well, especially for Spanish. We checked the data and found that the overall language model score for translated Spanish is better than reference, which means for Spanish, the fluency is not the big problem.

Semantic feature's performance is not stable from different language pairs. For C-E, it improves, but for S-E, it does. The reasons might be that Spanish is more like English, and the use of synonym does not occur much.

Also, experiments show that user information should be kept to make more confident evaluation.

## 5.3 Selection Experiments

Besides of evaluating the human correction quality by correlation, we also apply another selection experiment to see if there is a way that we can pick up good human corrections and feed them back to machine translation system.

### 5.3.1 Inner Selection

We use the combined features that perform best in previous experiment, which combines the score of sentence similarity, correction similarity and user quality. From figure 1&2, we can see that if we pick up the human correction with the highest score from each translation set, we can achieve comparable results as the human reference. Most important, the human corrections with score below 3 are total filtered out, which means that the worst human corrections are removed.

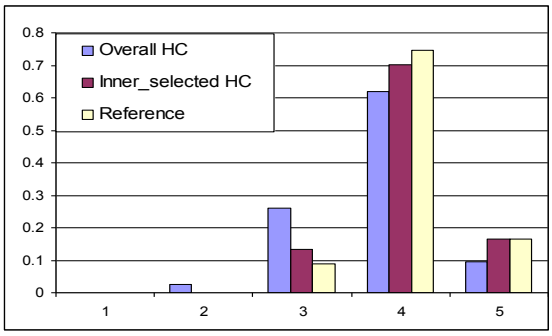


Figure1. Sentence quality distribution

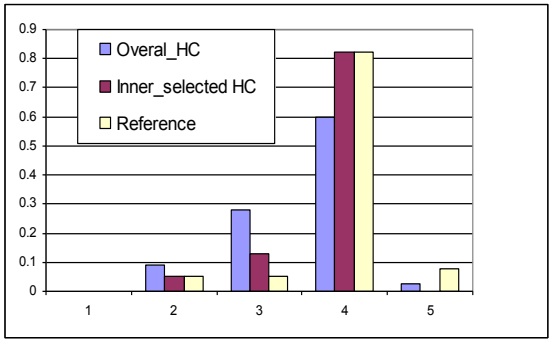


Figure2. Sentence quality distribution

### 5.3.2 Overall Selection

In this experiment, we only interested in the corrections with a human assessment above 4, which is good enough with the *reference quality*. Figure 3&4 shows that, the less corrections we selected, the more good corrections we get. Thus, we can easily set the threshold to return a subset of crowdsourcing data with higher qualities.

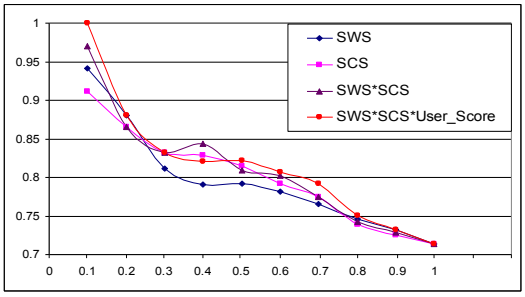


Figure3. Percentages of *reference quality* corrections in Chinese-English

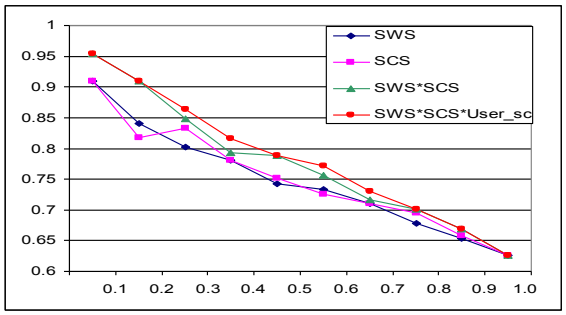


Figure4. Percentages of *reference quality* corrections in Spanish-English

## 6 Conclusions and Future Work

We evaluated the human correction qualities based on multiple corrections. In this way, we could cross validate the quality of a single correction. We investigated different features and compare their correlation to human assessment. We also tried to rank the quality of human corrections from good to bad, which enabled us to set a threshold to control the qualities of the human corrections.

**Acknowledgments**

We would like to thank other co-workers from IBM, including Ea-Ee Jan, Sasha Caskey, Hui Wan, Fei Huang and Jia Cui.

## References

K. Papineni, et al. BLEU: a method for automatic evaluation of machine translation. *IBM research division technical report, RC22176 (W0109-022), 2001.*

Brabham, D.C. 2008. Crowdsourcing as a model for problem solving: an introduction and cases. *In Convergence: International Journal of Research into New Media Technologies, 14, (1), 75-90*

R Snow, B O'Connor, D Jurafsky, AY. 2008. Cheap and fast---but is it good? evaluating non-expert annotations for natural language tasks. *Proceedings of EMNLP 2008, Honolulu*

C Callison-Burch. 2009. Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. *Proceeding of EMNLP 2009, Singapore*

M Marge, S Banerjee, A Rudnicky. 2010. Using the Amazon Mechanical Turk for transcription of spoken language. *Proceeding of ICASSP, March, 2010*

Enrique Amig´o, Jes´us Gim´enez, Julio Gonzalo, and Felisa Verdejo. 2009. The contribution of linguistic features to automatic machine translation evaluation. *In Proceedings of ACL 2009.*

M Gamon, A Aue, M Smets. 2005. Sentence-Level MT evaluation without reference translations: beyond language modeling. *Proceedings of EAMT, 2005*

Quirk, Christopher. 2004. Training a Sentence-Level Machine Translation Confidence Measure. In Proceedings of LREC 2004, pp 525-828.

Alex Kulesza and Stuart M. Shieber. 2004. A Learning Approach to Improving Sentence-Level MT Evaluation. *In Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*

Daniel Cer, Marie-Catherine de Marneffe, Daniel Jurafsky, and Christopher D. Manning. 2010. Parsing to Stanford Dependencies: Trade-offs between speed and accuracy. *In 7th International Conference on Language Resources and Evaluation (LREC 2010)*

D. Lin. 1998. An information-theoretic definition of similarity. *In Proceedings of the International Conference on Machine Learning, Madison, August.*

Jay J. Jiang and David W. Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *In Proceedings of International Conference Research on Computational Linguistics (ROCLING X), Taiwan*

P. Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pages 448–453, Montreal, August.*

Radu Soricut, A Echihabi. 2010. TrustRank: Inducing Trust in Automatic Translations via Ranking. *Proceedings of ACL 2010.*

Osamuyimen Stewart, David Lubensky, Juan M. Huerta . 2010. Crowdsourcing participation inequality: a SCOUT model for the enterprise domain *Proceedings of the ACM SIGKDD Workshop on Human Computation*