# Translation quality prediction using multiple automatic evaluation metrics

*Michael Paul*[†‡] *and Andrew Finch*[†‡] *and Eiichiro Sumita*[†‡]

† NICT Spoken Language Communication Group
‡ ATR Spoken Language Communication Research Labs
Hikaridai 2-2-2, Keihanna Science City, 619-0288 Kyoto
{Michael.Paul,Andrew.Finch,Eiichiro.Sumita}@nict.go.jp

## Abstract

This paper applies a machine learning method to predict human assessments of machine translation (MT) quality based on multiple automatic evaluation measures. Various automatic evaluation measures have been proposed, whereby each automatic metric focuses on different aspects of the translation output. However, none of these automatic metrics turned out to be satisfactory in judging the translation quality of a single translation. In order to tap the full potential of each metric and to predict human assessments more accurately, a supervised learning method is applied to learn discriminative models (*classifiers*) based on the results of multiple automatic evaluation metrics for a given translation. The effectiveness of the proposed method is evaluated for English translations in the travel domain.

## 1. Introduction

The evaluation of MT quality by humans is cost- and time-intensive. Various automatic evaluation measures have been proposed to make evaluations of MT outputs cheaper and faster. Recent evaluation campaigns on newswire[1] and travel data[2] investigated how good these evaluation metrics correlate with human judgments. The results showed that high correlations to human judges were obtained for some metrics when ranking MT system outputs on the document-level. However, each automatic metric focuses on different aspects of the translation output and its correlation towards human judges depends on the type of human assessment like *fluency* or *adequacy*. Moreover, none of the automatic metrics turned out to be satisfactory in predicting the translation quality of a single translation.

This paper applies a supervised learning method to predict human assessments based on the results of multiple automatic evaluation metrics for a given translation. The learned discriminative models boost the effects of each automatic evaluation metric by combining multiple indicators of translation quality automatically.

Section 2 describes the human and automatic evaluation metrics investigated in this paper. Section 3 introduces our proposed method and compares it to related research. The effectiveness of the proposed method is evaluated in Section 4 for English translations in the travel domain.

[1]NIST MT evaluations, http://www.nist.gov/speech/tests/mt
[2]IWSLT evaluations, http://www.slc.atr.jp/IWSLT2006

## 2. Assessment of Translation Quality

Various approaches on how to assess the qualilty of a translation has been proposed. In this paper, human assessments of translation quality with respect to the *acceptability*, the *fluency* and the *adequacy* of the translation are investigated. *Acceptability* judges how easy-to-understand the translation is [1]. *Fluency* indicates how the evaluation segment sounds to a native speaker of English. For *adequacy*, the evaluator was presented with the source language input as well as a "gold standard" translation and has to judge how much of the information from the original translation is expressed in the translation [2]. The *acceptability*, *fluency* and *adequacy* judgments consist of one of the grades listed in Table 1.

The high cost of such human evaluation metrics triggered a huge interest in the development of automatic evaluation metrics for machine translations. Table 2 introduces some metrics that are widely used in the MT research community.

## 3. Prediction of Human Assessments

Most of the previously proposed approaches to predict human assessments of translation quality utilize supervised learning methods like *decision trees*, *support vector machines*, or *perceptrons* to learn discriminative models that are able to come closer to human quality judgments. Such classifiers can be trained on a set of features extracted from human-evaluated MT system output.

The work described in [10] uses statistical measures to estimate confidence on the word/phrase level and gathers system-specific features about the translation process itself to train binary classifiers that distinguish between good and bad translations.

[11] utilizes multiple *edit-distance* features where combinations of lexical (stem, word, part-of-speech) and semantic (thesausus-based semantic class) matches are used to compare MT system outputs with reference translations and to approximate human scores of *acceptability*.

The approach proposed in this paper also utilizes a supervised learning method to predict human assessments of trans-

Table 1: Human assessment

| *acceptability* | |
|---|---|
| 4 | Perfect Translation |
| 3 | Good Translation |
| 2 | Fair Translation |
| 1 | Acceptable Translation |
| 0 | Nonsense |

| *fluency* | |
|---|---|
| 4 | Flawless English |
| 3 | Good English |
| 2 | Non-native English |
| 1 | Disfluent English |
| 0 | Incomprehensible |

| *adequacy* | |
|---|---|
| 4 | All Information |
| 3 | Most Information |
| 2 | Much Information |
| 1 | Little Information |
| 0 | None |

Table 2: Automatic evaluation metrics

| | |
|---|---|
| BLEU: | the geometric mean of n-gram precision of the system output with respect to reference translations. Scores range between 0 (worst) and 1 (best) [3] |
| NIST: | a variant of BLEU using the arithmetic mean of weighted n-gram precision values. Scores are positive with 0 being the worst possible [4] |
| METEOR: | calculates unigram overlaps between a translation and reference texts using various levels of matches (*exact*, *stem*, *synonym*). Scores range between 0 (worst) and 1 (best) [5] |
| GTM: | measures the similarity between texts by using a unigram-based F-measure. Scores range between 0 (worst) and 1 (best) [6] |
| WER: | *Word Error Rate*: the minimal edit distance between the system output and the closest reference translation divided by the number of words in the reference. Scores are positive with 0 being the best possible [7] |
| PER: | *Position independent WER*: a variant of WER that disregards word ordering [8] |
| TER: | *Translation Edit Rate*: a variant of WER that allows phrasal shifts [9] |

lation quality. In contrast to previous approaches, the feature set consists solely of multiple automatic evaluation scores, thus having the following advantages:

- *robustness*:
  The method does not depend on a specific MT system nor on the target language. It can be applied without modification to any translation or target language as long as reference translations are available.
- *reliability*:
  The automatic combination of multiple evaluation metrics boosts the effects of each metric and predicts human assessments more accurately.

The prediction method is divided into two phases: (1) the *learning phase* in which the classifier is trained on the feature set that is extracted from a database of human-evaluated MT system outputs and (2) the *application phase* in which the classifier is applied to unseen sentences and predicts a human score. For training, all automatic scores of the evaluation metrics described in Table 2 are calculated for each translation and stored in a database together with the respective human scores. The database was then randomly split into 10 subsets and the cross-validation technique [12], was applied for evaluation, i.e., for each subset $S_i$ ($i$=1,...,10), a classifier was trained on all remaining subsets ($\bigcup S_j$, $j \neq i$) and its performance was evaluated on all translations of $S_i$.

## 4. Evaluation

The evaluation of the proposed method is carried out using the *Basic Travel Expression Corpus* (BTEC) that contains tourism-related sentences similar to those usually found in phrase books for tourists going abroad [13]. In total, 3524 Japanese input sentences were translated by MT systems of various types[3] producing 82,406 English translations. 54,576 translations were annotated with human scores for *acceptability* and 36,302 translations were annotated with human scores for *adequacy*/*fluency*. The distribution of the human scores for the given translations is summarized in Table 3.

Table 3: Human score distribution

| human assessment | score 4 | 3 | 2 | 1 | 0 |
|---|---|---|---|---|---|
| *acceptability* | 16.3% | 11.5% | 14.5% | 14.7% | 43.0% |
| *adequacy* | 30.8% | 13.1% | 10.9% | 16.5% | 28.7% |
| *fluency* | 32.5% | 10.4% | 16.7% | 17.5% | 22.9% |

For the experiments described in this section, we utilized a standard implementation of decision trees [14] to learn discriminative models. In addition to the seven automatic evaluation scores listed in Table 2, the 4-gram and 5-gram precision scores calculated by the BLEU/NIST metrics were also utilized, resulting in a total of nine training features. The proposed method was applied to the classification tasks listed in Table 4 whereby the training sentences annotated with human scores in paranthesis were merged to decide on the binary-class ("+1" vs. "-1") assignments.

Table 4: Classification tasks

| type | task | classes |
|---|---|---|
| *multi-class* | 43210 | "4" vs. "3" vs. "2" vs. "1" vs. "0" |
| *binary-class* | 4_3210 | "+1" (4) vs. "-1" (3 or 2 or 1 or 0) |
| | 43_210 | "+1" (4 or 3) vs. "-1" (2 or 1 or 0) |
| | 432_10 | "+1" (4 or 3 or 2) vs. "-1" (1 or 0) |
| | 4321_0 | "+1" (4 or 3 or 2 or 1) vs. "-1" (0) |

### 4.1. Multi-Class Prediction

The *baseline* prediction of each multi-class task is defined as the selection of the most frequent class assigned in the database (cf. Table 3). Table 5 compares the performance of the proposed method to the baseline prediction, where the *accuracy* measure, i.e., the percentage of correctly classified

[3] Most of the translations were generated by statistical MT engines, but 5 example-based and 5 rule-based MT systems were also utilized.

Table 5: Accuracy of multi-class prediction

| **43210** | *acceptability* | *adequacy* | *fluency* |
|---|---|---|---|
| baseline | 43.0% | 30.8% | 32.5% |
| proposed | **67.2%** | **63.8%** | **60.6%** |
| (gain) | (+24.2%) | (+33.0%) | (28.1%) |

Table 6: Accuracy of prediction grades

| grade | *acceptability* | *adequacy* | *fluency* |
|---|---|---|---|
| 4 | **77.3%** | **86.2%** | **82.9%** |
| 3 | 49.3% | 35.2% | 32.4% |
| 2 | 43.6% | 30.1% | 41.0% |
| 1 | 29.1% | 36.8% | 37.7% |
| 0 | **89.0%** | **81.1%** | **73.7%** |

sentences, was used for evaluation. The results show that the proposed method outperforms the baseline method for all types of human assessment gaining 24-33% in accuracy. The highest prediction accuracy of 67% is achieved for *acceptability*, where 64% and 61% accuracy are achieved for *adequacy* and *fluency*, respectively.

Table 6 shows the prediction accuracies achieved for each grade. Very high accuracy figures are achieved for the highest grade (4) and the lowest grade (0) of all types of human assessment, but the system shows reduced performance on all medium grades. This tendency can also be found for human judges. Inter-evaluator agreement statistics of evaluation experiments have shown, that humans can easily identify *very good* and *very bad* translations, but judgments vary largely for translations of medium quality.

### 4.2. Binary-Class Prediction

The *baseline* prediction of each binary-class task is defined in analogy to the multi-class task, i.e., the maximal sum of all human scores grouped together within the respective binary classes. The accuracy figures of both methods are summarized in Table 7. The results show that the proposed method also outperforms the baseline system for all binary classes. However, due to differences in the score distributions for each classification type, the gains with respect to the baseline vary between the classifiation tasks. The accuracy of the proposed method for binary classification is 80-86%.

Similar to the multi-class task, the best system performance is achieved for *acceptability*, followed by *adequacy* and *fluency* for all binary classification tasks besides *4321_0*. The reason for this phenomenon is that the predictive power of a feature depends not only on the classification task, but also on the type of human assessment to be classified.

### 4.3. Feature Dependency

In order to get an idea of how the classification performance is effected by the respective features, we conducted two additional experiments: (1) train classifiers using only one feature (*feature only*) and (2) train classifiers on all features excluding a single feature (*w/o feature*). The feature dependencies for the multi-class prediction task are summarized in Table 8. The most *discriminative* feature is defined as one that achieves

Table 7: Accuracy of binary-class prediction

| **4_3210** | *acceptability* | *adequacy* | *fluency* |
|---|---|---|---|
| baseline | 83.7% | 69.2% | 67.5% |
| proposed | **91.2%** | **86.1%** | **82.7%** |
| (gain) | (+7.5%) | (+16.9%) | (+15.2%) |
| **43_210** | *acceptability* | *adequacy* | *fluency* |
| baseline | 72.1% | 56.0% | 57.0% |
| proposed | **85.4%** | **82.6%** | **78.8%** |
| (gain) | (+13.3%) | (+26.6%) | (+21.8%) |
| **432_10** | *acceptability* | *adequacy* | *fluency* |
| baseline | 57.7% | 54.8% | 59.7% |
| proposed | **80.3%** | **79.7%** | **76.9%** |
| (gain) | (+22.6%) | (+24.9%) | (+17.2%) |
| **4321_0** | *acceptability* | *adequacy* | *fluency* |
| baseline | 57.0% | 71.3% | 77.2% |
| proposed | **77.0%** | **79.7%** | **80.3%** |
| (gain) | (+20.0%) | (+8.4%) | (+3.1%) |

the highest accuracy when the classifier is trained on a single feature. The most *contributive* feature is defined as the one that obtains the lowest accuracy when omitted for the classifier training. Both types of features are highlighted in bold-face in the respective *feature only*, and *w/o feature* parts of Table 8.

For *acceptability* and *adequacy* judgments, the *METEOR* feature is the most important one, being both the most discriminative and the most contributive feature. For *fluency*, *METEOR* is also important, but less discriminative than the *4-gram* feature and less contributive than the *WER* feature. On the other hand, the *BLEU* feature is the least discriminative feature for all types of human assessment and does not contribute any gain for *acceptability*. Moreover, the least contributive features for *adequacy* are *TER* and *NIST*, where *TER* can also be omitted for the prediction of *fluency* grades.

These results complement findings of the recent evaluation campaigns [15] where the highest correlation between *adequacy*/*METEOR* and *fluency*/*BLEU* were obtained on the document-level. Whereas, the *METEOR* feature is also applicable at the sentence level, the *BLEU* metric is not helpful at all. One reason is that the *BLEU* metric assigns a score of 0 to all translations that do not match at least a sequence of 4 words in the reference translations. However, the *4-gram* precision scores calculated by the *BLEU* metrics proved to be applicable at the sentence level boosting the classification performance for *fluency*.

The effects of combining multiple features for the classifier training is summarized in Table 9. It lists the difference in accuracy performance between the proposed method using multiple features and the classifiers trained only on the most discriminative feature of the respective classification task. The results show that the proposed method outperforms all single-feature classifiers. The largest gain of up to 13% in accuracy is achieved for the *multi-class* prediction task whereas the combination of multiple features gains at least 2-4% in accuracy for the binary classification tasks.

Table 8: Feature dependency for multi-class prediction

| **all features** | *acceptability* | *adequacy* | *fluency* |
|---|---|---|---|
| *proposed* | **67.2%** | **63.8%** | **60.6%** |

| **feature only** | *acceptability* | *adequacy* | *fluency* |
|---|---|---|---|
| *BLEU* | 49.8% | 44.8% | 37.7% |
| *NIST* | 50.5% | 49.5% | 45.6% |
| *METEOR* | **54.2%** | **52.7%** | 48.3% |
| *GTM* | 49.9% | 50.7% | 43.9% |
| *WER* | 53.1% | 51.7% | 47.7% |
| *PER* | 51.9% | 51.8% | 46.0% |
| *TER* | 51.9% | 47.3% | 44.3% |
| *4gram* | 52.0% | 51.7% | **48.7%** |
| *5gram* | 50.6% | 51.3% | 41.4% |

| **w/o feature** | *acceptability* | *adequacy* | *fluency* |
|---|---|---|---|
| *BLEU* | 67.2% | 63.6% | 60.3% |
| *NIST* | 66.9% | 63.8% | 60.0% |
| *METEOR* | **65.2%** | **62.7%** | 59.6% |
| *GTM* | 66.1% | 63.3% | 60.2% |
| *WER* | 66.9% | 63.8% | **59.2%** |
| *PER* | 66.8% | 63.5% | 59.9% |
| *TER* | 66.8% | 63.8% | 60.6% |
| *4gram* | 66.6% | 63.4% | 60.0% |
| *5gram* | 66.2% | 63.2% | 59.9% |

Table 9: Effects of feature combination

| task | *acceptability* | *adequacy* | *fluency* |
|---|---|---|---|
| **43210** | +13.0% | +11.1% | +11.9% |
| **4_3210** | +2.1% | +2.7% | +2.7% |
| **43_210** | +3.6% | +3.0% | +3.3% |
| **432_10** | +4.0% | +3.9% | +2.8% |
| **4321_0** | +3.4% | +1.9% | +1.7% |

## 5. Conclusion

In this paper, we proposed a robust and reliable method to learn discriminative models to predict translation quality on the sentence level where the prediction is carried out by utilizing the results of multiple automatic evaluation metrics.

The effectiveness of the proposed method was verified for three types of human assessment of translation quality commonly used within the MT research community. Experiments on *multi-class* and *binary-class* prediction tasks showed that the combination of multiple evaluation metric features outperforms the baseline method that selects the most frequent class of the training data and the classsifiers trained on single features only, gaining up to 33% and 13% in prediction accuracy, respectively. The analysis of feature dependencies revealed, that the most important feature to predict the *acceptability* and *adequacy* of a translation is the *METEOR* metric, whereas *4-gram* precision and *WER* scores are helpful in predicting the *fluency* of a translation.

## 6. References


[1] E. Sumita, S. Yamada, K. Yamamoto, M. Paul, H. Kashioka, K. Ishikawa, and S. Shirai, "Solutions to problems inherent in spoken-language translation: The ATR-MATRIX approach," in *Proc. of the Machine Translation Summit VII*, Singapore, 1999, pp. 229–235.

[2] J. White, T. O'Connell, and F. O'Mara, "The ARPA MT evaluation methodologies: evolution, lessons, and future approaches," in *Proc of the AMTA*, 1994, pp. 193–205.

[3] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. of the 40th ACL*, Philadelphia, USA, 2002, pp. 311–318.

[4] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proc. of the HLT 2002*, San Diego, USA, 2002, pp. 257–258.

[5] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, 2005, pp. 65–72.

[6] J. Turian, L. Shen, and I. Melamed, "Evaluation of machine translation and its evaluation," in *Proc. of the MT Summit IX*, New Orleans, USA, 2003, pp. 386–393.

[7] S. Niessen, F. J. Och, G. Leusch, and H. Ney, "An evaluation tool for machine translation: Fast evaluation for machine translation research," in *Proc. of the 2nd LREC*, Athens, Greece, 2000, pp. 39–45.

[8] F. J. Och and H. Ney, "Statistical multi-source translation," in *Proc. of the MT Summit VIII*, Santiago de Compostella, Spain, 2001, pp. 253–258.

[9] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. of the AMTA*, Cambridge and USA, 2006, pp. 223–231.

[10] C. Quirk, "Training a sentence-level machine translation confidence measure," in *Proc. of 4th LREC*, Lisbon, Portugal, 2004, pp. 825–828.

[11] Y. Akiba, K. Imamura, and E. Sumita, "Using multiple edit distances to automatically rank machine translation output," in *Proc. of MT Summit VIII*, 2001, pp. 15–20.

[12] T. Mitchell, *Machine Learning*. New York, USA: The McGraw Hill Companies Inc., 1997.

[13] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," in *Proc. of the EUROSPEECH03*, Geneve, Switzerland, 2003, pp. 381–384.

[14] Rulequest, "Data mining tool c5.0," http://rulequest.com/see5-info.html, 2004.

[15] M. Paul, "Overview of the IWSLT Evaluation Campaign," in *Proc. of the IWSLT*, Japan, 2006, pp. 1–15.