International Journal of Computer Applications (0975 - 8887)
Volume 59 - No. 10, December 2012

Predicting Human Assessment of Machine Translation Quality by Combining Automatic Evaluation Metrics using Binary Classifiers

Michael Paul and Andrew Finch and Eiichiro Sumita
MASTAR Project
National Institute of Information and Communcations Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619–0289

ABSTRACT

This paper presents a method to predict human assessments of machine translation (MT) quality based on a combination of binary classifiers using a *coding matrix*. The multiclass categorization problem is reduced to a set of binary problems that are solved using standard classification learning algorithms trained on the results of multiple automatic evaluation metrics. Experimental results using a large-scale human-annotated evaluation corpus show that the decomposition into binary classifiers achieves higher classification accuracies than the multiclass categorization problem. In addition, the proposed method achieves a higher correlation with human judgments on the sentence level compared to standard automatic evaluation measures.

General Terms:

Machine Translation, Machine Learning, Machine Translation Evaluation

Keywords:

Evaluation Metric Combination, Human Assessment Prediction-ifx

1. INTRODUCTION

The evaluation of MT quality is a difficult task because there may exist many possible ways to translate a given source sentence. Moreover, the usability of a given translation depends on numerous factors such as the intended use of the translation, the characteristics of the MT software, and the nature of the translation process. Early attempts tried to manually produce numerical judgments of MT quality with respect to a set of reference translations [25]. Recent human assessment of MT quality has been carried out by either assigning a single grade on a scale of 5 or 7 specifying the fluency or adequacy of a given translation [19] or relatively ranking multiple translations of the same input to each other [7].
Although the human evaluation of MT output provides the most direct and reliable assessment, it is time consuming, costly and subjective, i.e., evaluation results might vary from person to person for the same translation output due to different backgrounds, bilingual experience, and inconsistent judgments caused by the high complexity of the multiclass grading task.
These drawbacks of human assessment schemes have encouraged many researchers to seek reliable methods for estimating such measures automatically. Various automatic evaluation measures have been proposed to make the evaluation of MT outputs cheaper and faster. However, automatic metrics have not yet proved able to consistently predict the usefulness of MT technologies. Each automatic metric focuses on different aspects of the translation output and its correlation with human judges depends largely on the type of human assessment.
Moreover, recent evaluation campaigns on newswire [19, 18] and travel data [17] investigated how well 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, none of the automatic metrics turned out to be satisfactory in predicting the translation quality of a single translation.
In order to overcome the above shortcomings of automatic evaluation metrics and combine their strengths, this paper investigates the usage of multiple evaluation metrics to predict human assessments of machine translation (MT) quality. The proposed method applies multiple automatic evaluation metrics, such as BLEU and METEOR, to a given translation task and uses the obtained sentence-level metric scores as features of the standard classification learning algorithms in order to predict the subjective grades assigned by humans for a given translation. The proposed method reduces the complexity of such multiclass categorization problems by (1) dividing the multiclass problem into binary classification tasks, (2) learning discriminative models based on features extracted from multiple automatic evaluation metric results, (3) producing binary indicators of translation quality on the sentence level, and (4) solving the multiclass classification problem by combining the results of the binary classifiers using a *coding matrix*. The main advantages of the proposed method are:

—the *reduction of classification ambiguity* due to the decomposition of a multiclass classification task into a set of binary classification problems;
—the *combination of multiple automatic evaluation metrics* to predict human judgments taking into account different aspects of translation quality.

The framework for reducing multiclass to binary classification and the combination of the binary results to solve the multiclass classification problem are described in Section 2. The human and automatic evaluation metrics investigated in this paper are described in Section 3. Section 4 gives a brief overview of related research on predicting human assessments and outlines the main differences of the proposed method. The proposed method is described in Section 5 and its effectiveness is evaluated in Section 6 for English translations of Chinese and Japanese source sentences in the travel domain.

2. MULTICLASS TO BINARY CLASSIFIER REDUCTION

Multiclass learning problems try to find an approximate definition of an unknown function whose range is a discrete set of values. Previous research on margin classifiers, investigated the feasibility of reducing multiclass categorization problems to multi-

ple binary problems that are solved using binary learning algorithms. The combination of the classifiers generated on the binary problems using error-correcting output codes is then used to solve the multiclass task [8]. The theoretical properties of combining binary classifiers to solve multiclass categorization problems are investigated in [Allwein et al. 2000] [3].
There are many ways in which a multiclass problem can be decomposed into a number of binary classification problems. The most well-known approaches are the *one-against-all* and *all-pairs*. In the *one-against-all* approach, a classifier for each of the classes is trained where all training examples that belong to that class are used as positive examples and all others as negative examples. In the *all-pairs* approach, classifiers are trained for each pair of classes whereby all training examples that do not belong to any of the classes in question are ignored [11]. Such decompositions of the multiclass problem can be represented by a *coding matrix* $\mathcal{M}$, where each class c of the multiclass problem is associated with a row of binary classifiers b. If k is the number of classes and l is the number of binary classification problems, the coding matrix is defined as:

$$\mathcal{M} = (m_{i,j})_{i=1,\ldots,k;j=1,\ldots,l}$$
$$m_{i,j} \in \{-1, 0, +1\},$$

If the training examples that belong to class c are considered as positive examples for a binary classifier b, then $m_{c,b}$=+1. Similarly, if $m_{c,b}$=–1 the training examples of class c are used as negative examples for the training of b. $m_{c,b}$=0 indicates that the respective training examples are not used for the training of classifier b [8, 3]. Examples of coding matrices for *one-against-all* and *all-pairs* (k=3, l=3) are given in Table 1.

Table 1. Coding Matrix Examples

one-against-all

	$c_1 \bullet c_{23}$	$c_2 \bullet c_{13}$	$c_3 \bullet c_{12}$
c_1	+1	–1	–1
c_2	–1	+1	–1
c_3	–1	–1	+1

all-pairs

	$c_1 \bullet c_2$	$c_1 \bullet c_3$	$c_2 \bullet c_3$
c_1	+1	+1	0
c_2	–1	0	+1
c_3	0	–1	–1

In this paper, the above coding matrix approach is applied to predict the outcomes of subjective evaluation metrics introduced in Section 3.

3. ASSESSMENT OF TRANSLATION QUALITY

Various approaches on how to assess the quality of a translation have been proposed. In this paper, human assessments of translation quality with respect to the *fluency*, the *adequacy* and the *acceptability* of the translation are investigated. *Fluency* indicates how natural the evaluation segment sounds to a native speaker of English. For *adequacy*, the evaluator is 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 [26]. *Acceptability* judges how easy to understand the translation is [23]. The *fluency*, *adequacy* and *acceptability* judgments consist of one of the grades listed in Table 2.
The high cost of such human evaluation metrics has triggered a huge interest in the development of automatic evaluation metrics for machine translation. Table 3 introduces some metrics that are widely used in the MT research community.

4. PREDICTION OF HUMAN ASSESSMENTS

Most of the previously proposed approaches to predict human assessments of translation quality utilize supervised learning methods such as *decision trees* (DT), *support vector machines* (SVM), or *perceptrons* to learn discriminative models that are able to come closer to human quality judgments. Such classifiers can

Table 2. Human Assessment

fluency	
5	Flawless English
4	Good English
3	Non-native English
2	Disfluent English
1	Incomprehensible

adequacy	
5	All Information
4	Most Information
3	Much Information
2	Little Information
1	None

acceptability	
5	Perfect Translation
4	Good Translation
3	Fair Translation
2	Acceptable Translation
1	Nonsense

Table 3. 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) [16]
NIST:	a variant of BLEU using the arithmetic mean of weighted n-gram precision values. Scores are positive with 0 being the worst possible [9]
METEOR:	calculates unigram overlaps between a translation and reference texts. For the experiments reported in this paper, only *exact matches* were taken into account. Scores range between 0 (worst) and 1 (best) [4]
GTM:	measures the similarity between texts by using a unigram-based F-measure. Scores range between 0 (worst) and 1 (best) [24]
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 [14]
PER:	*Position independent WER*: a variant of WER that disregards word ordering [15]
TER:	*Translation Edit Rate*: a variant of WER that allows phrasal shifts [22]

be trained on a set of features extracted from human-evaluated MT system outputs.
The work described in [Quirk 1994] [20] 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. Empirical thresholds on automatic evaluation scores are utilized to distinguish between good and bad translations. [Quirk 1994] also investigates the feasibility of various learning approaches for the multiclass classification problem for a very small data set in the domain of technical documentation. [Akiba et al. 2001] [2] utilized DT classifiers trained on multiple *edit-distance* features where combinations of lexical (stem, word, part-of-speech) and semantic (thesaurus-based semantic class) matches were used to compare MT system outputs with reference translations and to approximate human scores of *acceptability* directly. [Kulesza and Shieber 2004] [13] trained a binary SVM classifier based on automatic scoring features in order to distinguish between "human-produced" and "machine-generated" translations of newswire data instead of predicting human judgments directly.
The proposed approach also utilizes a supervised learning method to predict human assessments of translation quality, but differs in the following two aspects:

(1) *Reduction of Classification Ambiguity*:
The decomposition of a multiclass classification task into a set of binary classification problems reduces the complexity of the learning task, resulting in higher classification accuracy.

(2) *Feature Set*:
Classifiers are trained on the results of multiple automatic

evaluation metrics, thus taking into account different aspects of translation quality addressed by each of the metrics. The method does not depend on a specific MT system nor on the target language.

5. HUMAN ASSESSMENT PREDICTION BY BINARY CLASSIFIER COMBINATION

The proposed prediction method is divided into three phases: (1) a *learning phase* in which binary classifiers are trained on the feature set that is extracted from a database of human and machine-evaluated MT system outputs, (2) a *decomposition phase* in which the optimal set of binary classifiers that maximizes the classification accuracy of the recombination step on a development set is selected, (3) an *application phase* in which the binary classifiers are applied to unseen sentences, and the results of the binary classifiers are combined using the optimized coding matrix to predict a human score. A flow chart summarizing the major processing steps of each phase is given in Figure 1.

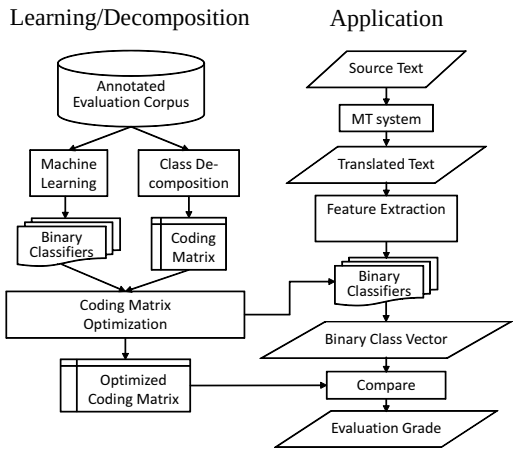

Fig. 1. Flow Chart of the Proposed Prediction Method

5.1 Learning Phase

Discriminative models for the multiclass and binary classification problem are obtained by using standard learning algorithms. The proposed method is not limited to a specific classification learning method. For the experiments described in Section 6, a standard implementation of decision trees [21] was utilized.
The feature set consists of the scores of the seven automatic evaluation metrics listed in Table 3. All automatic evaluation metrics were applied to the input data sets consisting of English MT outputs whose translation quality was manually assessed by humans using the metrics introduced in Section 3. In addition to the metric scores, metric-internal features, like *ngram-precision* scores, and *length ratios* between references and MT outputs, were also utilized, resulting in a total of 54 training features.

5.2 Decomposition Phase

For the experiments described in Section 6, both the *one-against-all* and the *all-pairs* binary classifiers described in Section 2 were utilized . In addition, *boundary* classifiers were trained on the whole training set. In this case, all training examples annotated with a class better than the class in question were used as positive examples, and all other training examples as negative examples. Table 4 lists the 17 binary classification problems that were utilized to decompose the human assessment problems introduced in Section 3.

Table 4. Decomposition of Human Assessment of Translation Quality

type	binary classifier
one-against-all	5, 4, 3, 2, 1
all-pairs	5_4, 5_3, 5_2, 5_1, 4_3, 4_2, 4_1, 3_2, 3_1, 2_1
boundary	54_321, 543_21

In order to identify the optimal coding matrix for the respective tasks, the binary classifiers were first ordered according to their classification accuracy on the development set. In the second step, the multiclass performance was evaluated iteratively, where the worst performing binary classifier was omitted from the coding matrix after each iteration. Finally, the coding matrix achieving the best classification accuracy for the multiclass task was used for the evaluation of the test set.

5.3 Application Phase

Given an input example, all binary classifiers are applied once for each column of the coding matrix resulting in a vector v of l binary classification results. The multiclass label is predicted as the label c for which the respective row r of $\mathcal{M}$ is "closest". In [Allwein et al. 2000] [3], the distance between r and v is calculated by (a) a generalized *Hamming distance* that counts the number of positions for which the corresponding vectors are different and (b) a *loss-based decoding* that takes into account the magnitude of the binary classifier scores. For the experiments described in Section 6, the Hamming-distance approach was adopted.

Table 5. Coding Matrix Application
$v = (+1, +1, -1)$

type	multiclass	distance	selection
one-against-all	c_1	1	c_1 or c_2
	c_2	1	
	c_3	3	
all-pairs	c_1	1	c_1
	c_2	3	
	c_3	2	

An example for the distance calculation is given in Table 5. Let's assume that the application of the three binary classifiers listed in Table 1 results in the classification vector $v = (+1, +1, -1)$ for a given input. Using the *one-against-all* coding matrix, the minimal distance for v is 1 for both matrix rows, c_1 and c_2. In the case of a draw, the priority order of binary classifiers obtained on the development set is used to identify the more reliable row by recursively recalculating the distance for a subset of classifiers with the less accurate one removed. For the *all-pairs* coding matrix, class c_1 would be selected due to its lesser distance.

6. EVALUATION

The evaluation of the proposed method was carried out using the *Basic Travel Expression Corpus* (BTEC). This contains tourism-related sentences similar to those usually found in phrase books for tourists going abroad [12]. In total, 3,524 Japanese input sentences were translated by MT systems of various types[1], producing 82,406 English translations. 54,576 translations were an-

[1]Most of the translations were generated by statistical MT engines, but 5 example-based and 5 rule-based MT systems were also utilized. These engines were state-of-the-art MT engines. Some participated in the IWSLT evaluation campaign series and some were in-house MT engines.

notated 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 Figure 2. Where multiple human judgments were assigned to a single translation output, the median of the respective human scores was used in the experiments.

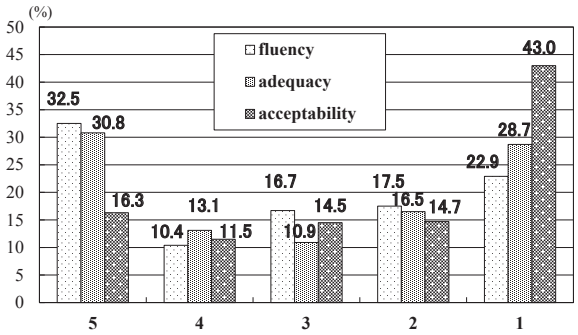

Fig. 2. Human Score Distribution

The annotated corpus was split into three data sets: (1) the training set (*train*) consisting of 25,988 translations for *adequacy/fluency* and 49,516 MT outputs for *acceptability*, (2) the development set (*devset*) consisting of 2,024 sentences (4 MT outputs for each of 506 input sentences) for all three metrics, and (3) the evaluation set (*testset*) taken from the IWSLT evaluation campaign [1] (CSTAR03 data set, 506 input sentences, up to 16 reference translations per sentence). For *fluency* and *adequacy*, 7,590 test sentences with 15 MT outputs for each were available. For *acceptability*, 3,036 sentences with 6 MT outputs for each were used for evaluation.

6.1 Coding Matrix Optimization

Figure 3 summarizes the iterative evaluation of the binary classification combination using the $devset$ as described in Section 5.2. Starting with the complete coding matrix (*ALL*), the worst performing binary classifier, i.e., the classifier that achieved the lowest accuracy when evaluated in isolation, is omitted in the next iteration. This iterative elimination process continues until only a single binary classifier is left. The dash square indicates the subset of binary classifiers selected for the coding matrix utilized for the test set evaluation.
For example, in the case of the *fluency* evaluation task, the binary classifier "3_2" achieved the lowest classification accuracy score and therefore is omitted in the first iteration. The re-evaluation of the reduced coding matrix after each iteration revealed that the overall system performance does not change until iteration 7, where the omission of the binary classifier "5_4" leads to an improvement in accuracy. The highest accuracy scores is achieved in iteration 14. Therefore, a coding matrix taking into account the binary classifiers "5_1", "5_3", "2", and "3" is applied to the test set evaluation of the *fluency* task.

6.2 Classification Accuracy

The baseline of the multiclass classification task was defined as the class most frequently occurring in the training data set. Table 6 summarizes the baseline performance for all three subjective evaluation metrics.

Table 6. Baseline Accuracy ($testset$)

fluency	*adequacy*	*acceptability*
32.5%	**30.8%**	**43.0%**

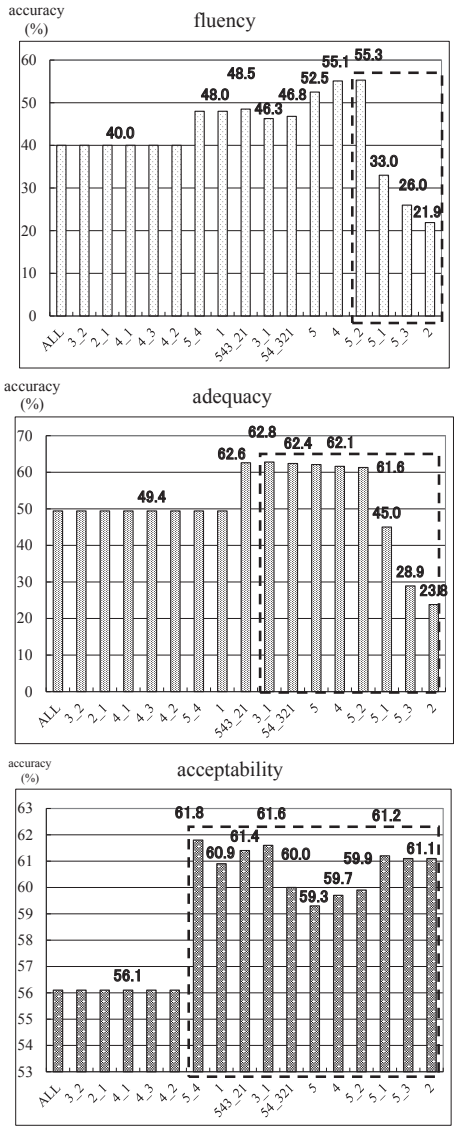

Fig. 3. Coding Matrix Optimization ($devset$)

The classification accuracies of the multiclass task, i.e. the multiclass classifier learned directly from the training set, and the binary classifier performance is summarized in Table 7. The results show that the learning approach outperforms the baseline of the multiclass classification task for all three metrics gaining 16.7% for *fluency*, 25.2% for *adequacy* and 18.1% for *acceptability*.

Table 7. Multiclass Prediction Accuracy ($testset$)

fluency	*adequacy*	*acceptability*
49.2%	**56.0%**	**61.1%**

Moreover, the accuracy figures of the binary classifiers are summarized in Figure 4. The performance of the binary classifiers varies widely, depending on the classification task as well as the evaluation metric. Accuracies of 78%-94% were achieved for the *one-against-all* classifiers, 76%-83% for the *boundary* classifiers, and 54%-92% for the *all-pairs* classifiers. However, the majority of the binary classifiers achieved a higher accuracy than the multiclass classifier for all evaluation metrics.
The proposed method combines the binary classifiers according to the optimized coding-matrix. The results are shown in

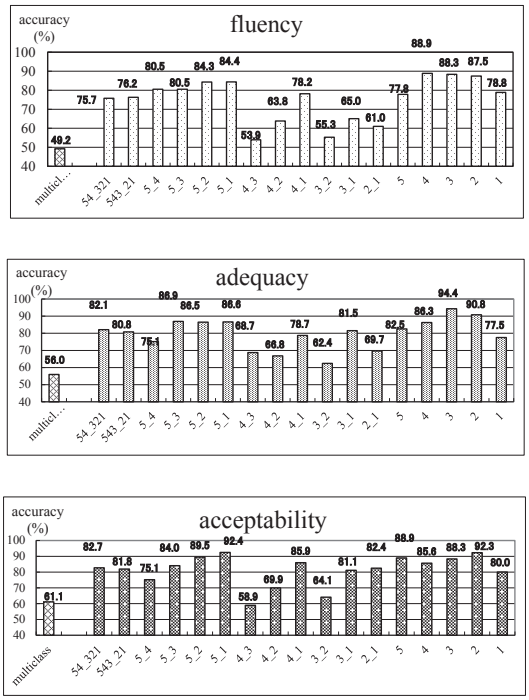

Fig. 4. Classifier Accuracy ($testset$**)**

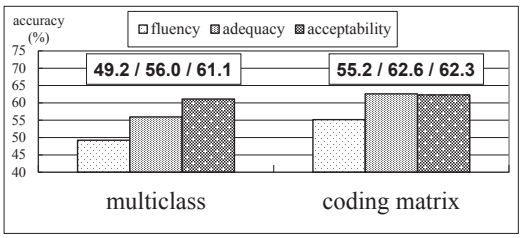

Fig. 5. Classifier Combination Accuracy ($testset$**)**

Figure 5. The classification accuracy of the proposed method is 55.2% for *fluency*, 62.6% for *adequacy* and 62.3% for *acceptability*. Thus, the proposed method outperforms the baseline as well as the multiclass classification task for all subjective evaluation metrics achieving a gain of 22.7% / 6.0% in *fluency*, 32.2% / 6.6% in *adequacy* and 19.3% / 1.2% in *acceptability* compared to the baseline / multiclass performance, respectively.

6.3 Correlation to Human Assessments

In order to investigate the correlation of the proposed metrics towards human judgments on the sentence level, the Spearman rank correlation coefficient for the obtained results was calculated. In addition, the multiclass classifier and the automatic evaluation metrics listed in Table 3 were used to rank the test sentences and calculate its Spearman rank correlation with human assessments.

The correlation coefficients are summarized in Figure 6. The results show that the proposed method outperforms all other metrics, achieving correlation coefficients of 0.632 / 0.759 / 0.769 for *fluency* / *adequacy* / *acceptability* where all score differences were statistical significant at the 95% level. Concerning the automatic evaluation metrics, METEOR achieved the highest correlation towards human assessment on the sentence level for all three subjective evaluation metrics. The correlation of the remaining automatic metrics is considerably lower and depends largely on the type of human assessment.

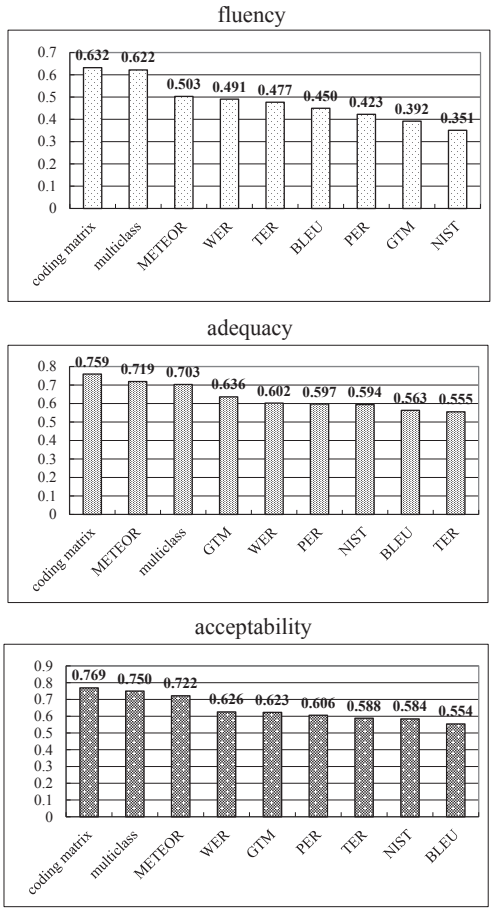

Fig. 6. Correlation with Human Assessments ($testset$**)**

6.4 Upper Bound

In order to get an idea about the potential of the proposed method, its upper bound was simulated by randomly adjusting the prediction result of each binary classifier to achieve a certain classification accuracy, and applied the coding matrix approach to the set of binary classifiers having the same classification accuracy. Figure 7 shows the upper boundary for classification accuracies between 60% and 100%, whereby the optimized coding matrix of the experiments described in Section 6.2 were used for *fluency*, *adequacy* and *acceptability*, respectively. The *all_binary* result shows the performance when the baseline coding matrix using all 17 binary classifiers is applied.

The results show that for each metrics the multiclass classification task performance is almost linearly related to the performance of the binary classifiers and that improving the accuracy of the binary classifiers will result in a better overall performance.

Two potential improvements of the proposed method to be investigated in the near future are (1) additional features that help to classify the given task more accurately, and (2) the automatic learning of the optimal combination of binary classifiers with respect to the overall system performance.

7. DISCUSSION

The guiding principle behind the proposed method is to use information on human annotations in order to be able to more closely approximate the human grading process. The price for this is that the method relies on human gradings of machine translation quality to annotate the training corpus used for the learning of the binary classifiers. However, to counter the high

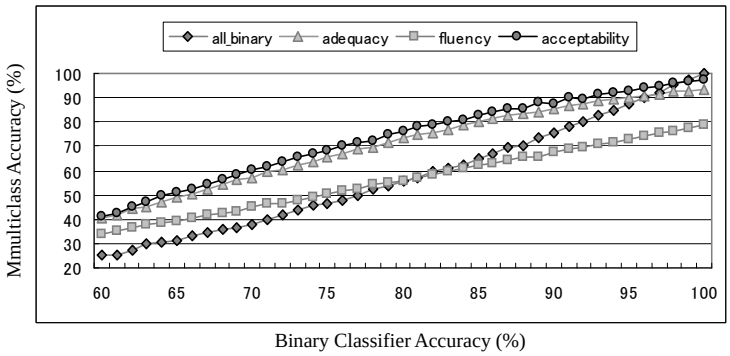

Fig. 7. Upper Boundary of Reducing Multiclass to Binary Classifier ($testset$)

costs in human assessment of MT outputs, new possibilities are offered by the advent of crowdsourcing services such as Amazon's Mechanical Turk[2] and CrowdFlower[3], which in recent years have attracted a lot of attention both from industry and academia as a means for collecting data for human language technologies at low cost [7, 5]. Having collected a certain amount of judgments for a specific translation task, the human data annotations and the derived prediction models can be re-used to judge the quality of translations of unseen sentences from the same translation domain without the need of humans in the loop. In addition, looking at the results of recent shared MT evaluation tasks, quite different score ranges were obtained for automatic evaluation metrics depending on (a) the domain from which the input sentences were taken, (b) the language pairs involved, and (c) the amount of reference translations used to calculate the automatic evaluation scores which gives rise to concerns on the scalability and robustness of the proposed method. However, score distribution transformation techniques such as *z-transform*[4], can be used to compare different metric scores of the same domain [17], but also to adapt already collected human annotated gradings to evaluation tasks having different score distributions. Hence, the proposed method of combining automatic evaluation metrics to predict human assessment of translation quality forms a general framework that can be applied across languages and translation domains.

8. CONCLUSION

In this paper, a robust and reliable method to learn discriminative models based on the results of multiple automatic evaluation metrics was proposed to predict translation quality at the sentence level. The prediction is carried out by reducing the multiclass classification problem to a set of binary classification tasks and combining the respective results using a coding matrix in order to predict the multiclass label for a given input sentence.
The effectiveness of the proposed method was verified using three types of human assessment of translation quality commonly used within the MT research community. The experiments showed that the proposed method outperforms a baseline method that selects the most frequent class contained in the training set and a standard multiclass classification model (decision tree) that learns its discriminative model directly from the training corpus. The proposed method achieved a gain of 22.7% / 6.0% in *fluency*, 32.2% / 6.6% in *adequacy* and 19.3% / 1.2% in *acceptability* compared to the baseline / multiclass performance, respectively. Moreover, the proposed metric achieved high correlation to human judgments at the sentence level, outperforming not only the multiclass approach but also all of the automatic scoring metrics utilized.
Future extensions of the proposed method will investigate the use of additional features, such as the confidence estimation features [6] and additional evaluation metrics like IMPACT [10] that obtained high correlations at sentence-level in recent MT evaluation campaigns. This is expected to improve the performance of the binary classifiers and boost the overall performance further.

[2]http://www.mturk.com
[3]http://crowdflower.com
[4]A transformation into score distributions with zero mean and unit variance.

9. REFERENCES

[1] Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii. Overview of the IWSLT04 evaluation campaign. In *Proc. of the International Workshop on Spoken Language Translation*, pages 1–12, Kyoto, Japan, 2004.

[2] Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita. Using multiple edit distances to automatically rank machine translation output. In *Proc. of MT Summit VIII*, pages 15–20, 2001.

[3] Erin Allwein, Robert Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

[4] Satanjeev Banerjee and Alon 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*, pages 65–72, Ann Arbor, Michigan, 2005.

[5] Luisa Bentivogli, Marcello Federico, Giovanni Moretti, and Michael Paul. Getting Expert Quality from the Crowd for MT Evaluation. In *Proceedings of the MT Summmit XIII*, pages 521–528, Xiamen, China, 2011.

[6] John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence estimation for SMT. In *Final Report of the JHU Summer Workshop*, 2003.

[7] Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. Findings of the 2010 joint workshop on smt and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on SMT and MetricsMATR*, pages 17–53, Uppsala, Sweden, 2010. Association for Computational Linguistics.

[8] Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[9] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In

Proc. of the HLT 2002, pages 257–258, San Diego, USA, 2002.

[10] Hiroshi Echizen-ya and Kenji Araki. Automatic evaluation of machine translation based on recursive acquisition of an intuitive common parts continuum. In *Proc. of the MT SUMMIT XI*, pages 151–158, Copenhagen, Denmark, 2007.

[11] Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, 26(2):451–471, 1998.

[12] Genichiro Kikui, Seiichi Yamamoto, Toshiyuki Takezawa, and Eiichiro Sumita. Comparative study on corpora for speech translation. *IEEE Transactions on Audio, Speech, Language Processing*, 14(5):1674–1682, 2006.

[13] Alex Kulesza and Stuart M. Shieber. A learning approach to improving sentence-level MT evaluation. In *Proc. of the TMI04*, USA, 2004.

[14] Sonja Niessen, Franz J. Och, Gregor Leusch, and Hermann Ney. An evaluation tool for machine translation: Fast evaluation for machine translation research. In *Proc. of the 2nd LREC*, pages 39–45, Athens, Greece, 2000.

[15] Franz J. Och and Hermann Ney. Statistical multi-source translation. In *Proc. of the MT Summit VIII*, pages 253–258, Santiago de Compostella, Spain, 2001.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th ACL*, pages 311–318, Philadelphia, USA, 2002.

[17] Michael Paul, Marcello Federico, and Sebastian Stücker. Overview of the IWSLT 2010 Evaluation Campaign. In *Proc. of IWSLT*, pages 3–27, Paris, France, 2010.

[18] Mark Przybocki and Kay Peterson. NIST Open Machine Translation Evaluation. http://www.nist.gov/speech/tests/mt, 2009.

[19] Mark Przybocki, Kay Peterson, and Sebastien Bronsart. Metrics for MAchine TRanslation Challenge (MetricsMATR08). http://nist.gov/speech/tests/metricsmatr/2008/results, 2008.

[20] Christopher B. Quirk. Training a sentence-level machine translation confidence measure. In *Proc. of 4th LREC*, pages 825–828, Portugal, 2004.

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

[22] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proc. of the AMTA*, pages 223–231, Cambridge and USA, 2006.

[23] Eiichiro Sumita, Setsuo Yamada, Kazuhide Yamamoto, Michael Paul, Hideki Kashioka, Kai Ishikawa, and Satoshi Shirai. Solutions to problems inherent in spoken-language translation: The ATR-MATRIX approach. In *Proc. of the MT Summit VII*, pages 229–235, Singapore, 1999.

[24] Joseph Turian, Luke Shen, and I. Melamed. Evaluation of machine translation and its evaluation. In *Proc. of the MT Summit IX*, pages 386–393, New Orleans, USA, 2003.

[25] John White, Theresa O'Connell, and Lynn Carlson. Evaluation of machine translation. In *Proc. of the Human Language Technology Workshop (ARPA)*, pages 206–210, 1993.

[26] John White, Theresa O'Connell, and Francis O'Mara. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Proc of the AMTA*, pages 193–205, 1994.