Reducing Human Assessment of Machine Translation Quality to Binary Classifiers

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 presents a method to predict human assessments of machine translation (MT) quality based on the 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.

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 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, each automatic metric focuses on different aspects of the translation output and its correlation towards human judges depends on the type of human assessment (for example *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 presents a method to predict human assessments of machine translation (MT) quality based on the combination of binary classifiers. The multiclass categorization problem is reduced to a set of binary problems that are solved using standard classification learning algorithms. Binary classifiers are trained on features of multiple automatic evaluation metrics, such as BLEU and METEOR. The learned discriminative models are applied sentence-wise to MT outputs producing binary indicators of translation quality on the sentence-level. The multiclass classification problem is then solved by combining the results of the binary classifiers using a *coding matrix*.

The human and automatic evaluation metrics investigated in this paper are described in Section 2. Section 3 gives a brief overview on related research on predicting human assessments and outlines the main differences to the proposed method. Section 4 outlines the

[1] NIST MT evaluations, http://www.nist.gov/speech/tests/mt

[2] IWSLT evaluations, http://www.slc.atr.jp/IWSLT2006

Table 1: Human Assessment

fluency		*adequacy*		*acceptability*	
5	Flawless English	5	All Information	5	Perfect Translation
4	Good English	4	Most Information	4	Good Translation
3	Non-native English	3	Much Information	3	Fair Translation
2	Disfluent English	2	Little Information	2	Acceptable Translation
1	Incomprehensible	1	None	1	Nonsense

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) (Papineni et al., 2002)
NIST:	a variant of BLEU using the arithmetic mean of weighted n-gram precision values. Scores are positive with 0 being the worst possible (Doddington, 2002)
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) (Banerjee and Lavie, 2005)
GTM:	measures the similarity between texts by using a unigram-based F-measure. Scores range between 0 (worst) and 1 (best) (Turian et al., 2003)
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 (Niessen et al., 2000)
PER:	*Position independent WER*: a variant of WER that disregards word ordering (Och and Ney, 2001)
TER:	*Translation Edit Rate*: a variant of WER that allows phrasal shifts (Snover et al., 2006)

proposed method. The framework of reducing multiclass to binary classification and the combination of the binary results to solve the multiclass classification problem are described in detail. The effectiveness of the proposed method is evaluated in Section 5 for English translations of Chinese and Japanese source sentences in the travel domain.

2 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 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 (White et al., 1994). *Acceptability* judges how easy-to-understand the translation is (Sumita et al., 1999). The *fluency*, *adequacy* and *acceptability* judgments consist of one of the grades listed in Table 1.

The high cost of such human evaluation metrics has triggered a huge interest in the development of automatic evaluation metrics for machine translation. 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* (DT), *support vector machines* (SVM), 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 outputs.

The work described in (Quirk, 2004) 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. He 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) utilized DT classifiers trained on multiple *edit-distance* features where combinations of lexical (stem, word, part-of-speech) and semantic (thesausus-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) 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 approach proposed in this paper also utilizes a supervised learning method to predict human assessments of translation quality, but differs in the following two aspects:

(1) *Reduction of Classification Perplexity*:
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 (see Table 2) 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. It can be applied without modification to any translation or target language as long as reference translations are available.

4 Human Assessment Prediction based on 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.

4.1 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 5, we utilized a standard implementation of decision trees (Rulequest, 2004).

The feature set consists of the scores of the seven automatic evaluation metrics listed in Table 2. 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 2. In addition to the metric scores, metric-internal features, like *ngram-precision* scores, *length ratios* between references and MT outputs, etc. were also utilized, resulting in a total of 54 training features.

4.2 Decomposition Phase

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 (Hastie and Tibshirani, 1998).

Such decompositions of the multiclass problem can be represented by a *coding matrix* $\mathcal{M}$

where each class c of the multiclass problem is associate 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:

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

$$m_{i,j} \in \{-1,\ 0,\ +1\},$$

where k is the number of classes and l is the number of binary classification problems. If the training examples that belong to class c are considered as positive examples for a binary classifier b, then $m_{c,b}$=+1. Similarily, 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 (Dietterich and Bakiri, 1995; Allwein et al., 2000). Examples of coding matrices for *one-against-all* and *all-pairs* (k=3, l=3) are given in Table 3.

Table 3: 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

For the experiments described in Section 5, we utilized both *one-against-all* and *all-pairs* binary classifiers. 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 2.

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. The optimized coding matrix reflects the standard bias-variance trade-off balancing the discriminative power and the reliability of the binary classifier combination.

Table 4: Decomposition of Human Assessment of Translaton 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

4.3 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), 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 5, we adopted the Hamming-distance approach.

An example for the distance calculation is given in Table 5. Lets assume that the application of the three binary classifiers listed in Table 3 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 case of a draw, the priority order of binary classifiers obtained on the development set is used to identify the more reliable row. For the *all-pairs* coding matrix, class c_1 would be selected due to its lesser distance.

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	

5 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 (Kikui et al., 2003). In total, 3,524 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 Figure 1. In case multiple human judgments were assigned to a single translation output, the median of the respective human scores was used in our experiments.

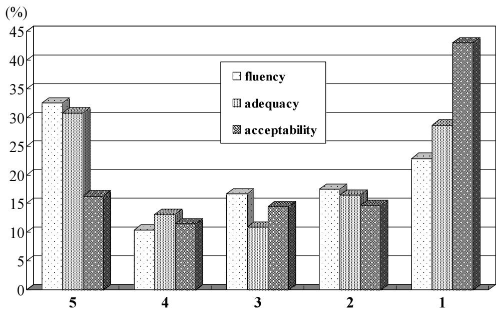

Figure 1: Human Score Distribution

The annotated corpus was split into three data sets: (1) the *training set* consisting of 25,988 translations for *adequacy/fluency* and 49,516 MT outputs for *acceptability*, (2) the *development set* consisted of 2,024 sentences (4 MT outputs for each of 506 input sentences) for all three metrics, and (3) the *test set* taken from the IWSLT evaluation campaign (CSTAR03 data set, 506 input sentences). 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.

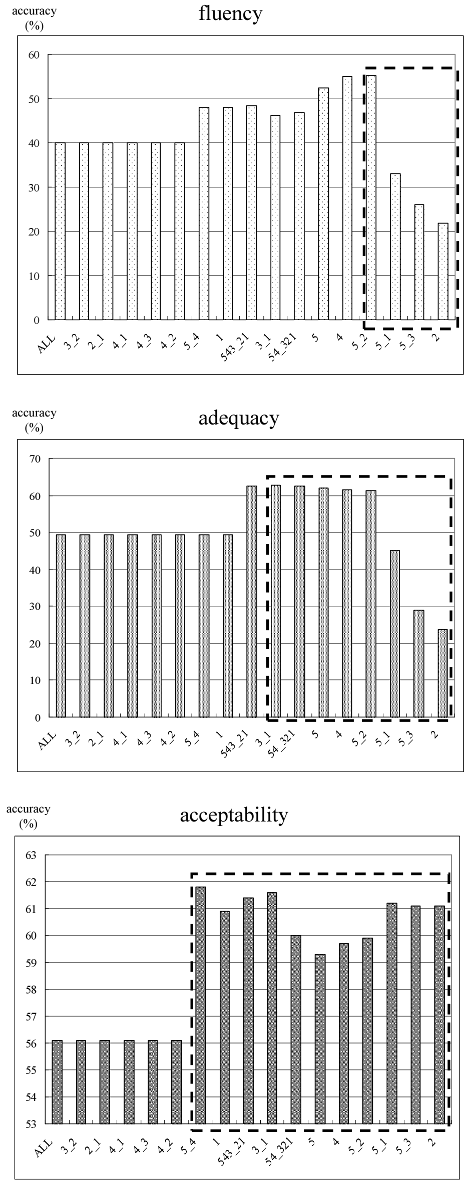

Figure 2: Coding Matrix Optimization

5.1 Coding Matrix Optimization

Figure 2 summarizes the iterative evaluation of the binary classification combination us-

[3] 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.

ing the development set as described in Section 4.2. Starting with the complete coding matrix (ALL), the worst performing binary classification is omitted in the next iteration. The dashed square indicates the subset of binary classifiers selected for the coding matrix utilized for the test set evaluation.

5.2 Classification Accuracy

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

Table 6: Baseline Accuracy

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

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 Figure 3. The results show that the learning approach outperforms the baseline of the multiclass classification task for all three metrics gaining 16.7% for *fluency*, 26.8% for *adequacy* and 18.1% for *acceptability*.

Moreover, the performance of the binary classifiers varies widely, depending on the classification task as well as the evaluation metric. Accuracies of 80%-90% were achieved for the *all-against-one* classifiers, 75%-81% for the *boundary* classifiers, and 55%-91% for the *all-pairs* classifiers.

The proposed method combines the binary classifiers according to the optimized coding-matrix. The results are shown in Figure 4. 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*, 31.5% / 6.6% in *adequacy* and 19.3% / 1.2% in *acceptability* compared to the baseline / multiclass performance, respectively.

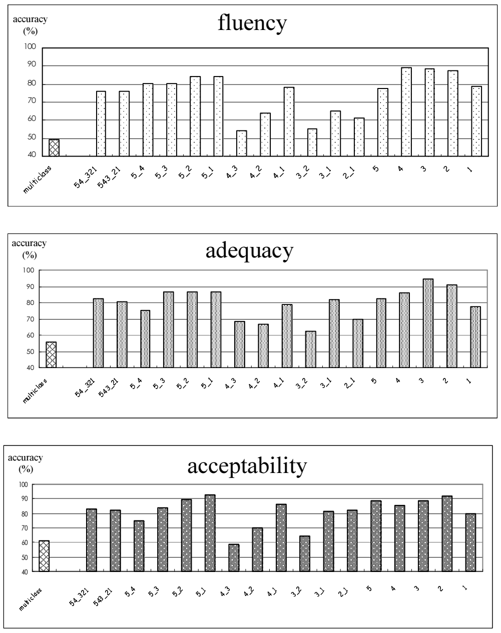

Figure 3: Classifier Accuracy

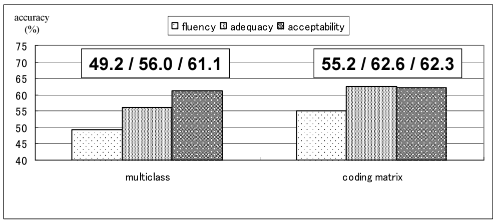

Figure 4: Classifier Combination Accuracy

5.3 Correlation to Human Assessments

In order to investigate the correlation of the proposed metrics towards human judgments on the sentence-level, we calculated the Spearman rank correlation coefficient for the obtained results. In addition, we used the multiclass classifier and the automatic evaluation metrics listed in Table 2 to rank the test sentences and calculate its Spearman rank correlation towards human assessments. The correlation coefficients are summarized in Figure 5.

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*, respectively. Concerning the automatic evaluation metrics, METEOR achieved the high-

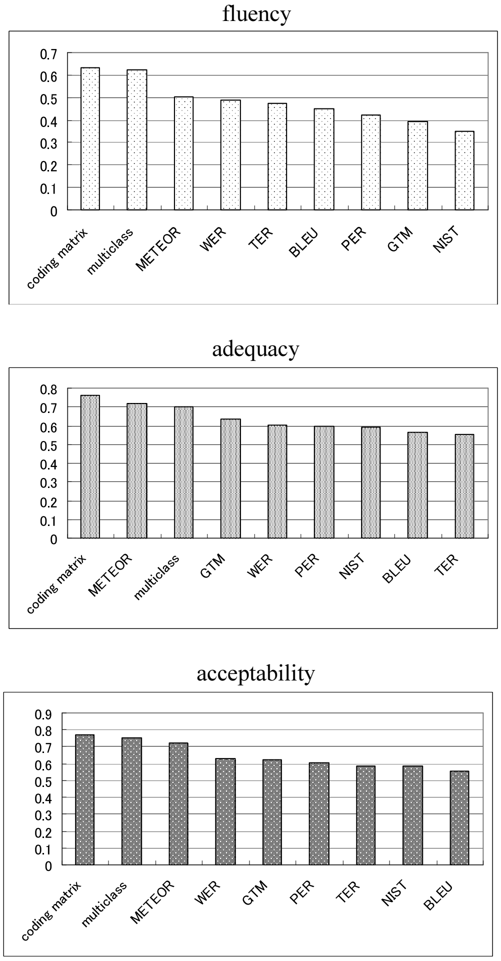

Figure 5: Correlation with Human Assessments

est correlation towards human assessment on 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.

5.4 Upper Bound

In order to get an idea about the potential of the proposed method, we simluated the upper bound of the method 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 6 shows the upper boundary of the proposed method for classification accuracies between 60% and 100% whereby the respective optimized coding matrix of the experiments described in Section 5.2 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, that we would like to investigate in the near future, are (1) additional features that help to classify the given task more acurately, and (2) the automatic learning of the optimal combination of binary classifiers with respect to the overall system performance.

6 Conclusion

In this paper, we proposed a robust and reliable method to learn discriminative models based on the results of multiple automatic evaluation metrics 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*, 31.5% / 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

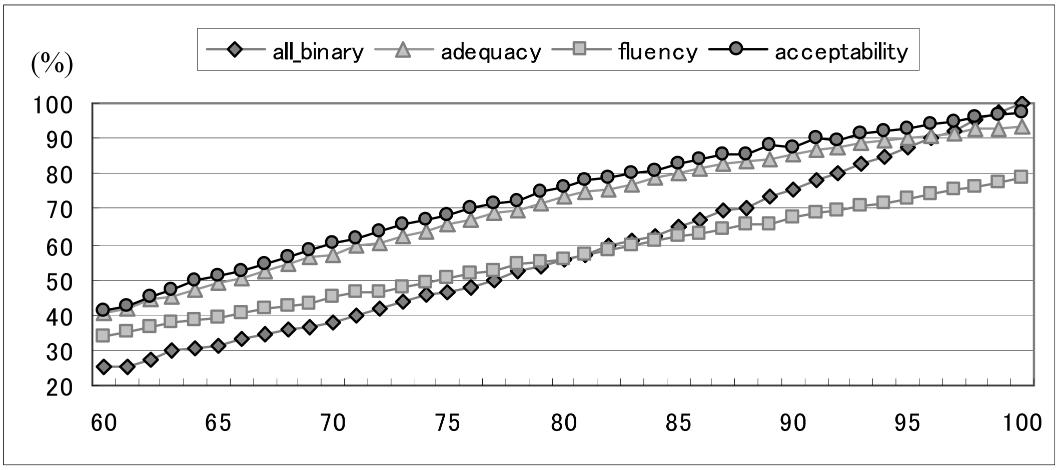

Figure 6: Upper Boundary of Reducing Multiclass to Binary Classifier

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 proposed in (Blatz et al., 2003) or the recently proposed source language features for MT evaluation in (Liu and Gildea, 2007). We would expect this to improve the performance of the binary classifiers and boost the overall performance further.

References

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

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

Satanjeev Banerjee and Alon Lavie. 2005. 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.

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

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

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. of the HLT 2002*, pages 257–258, San Diego, USA.

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

Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *Proc. of the EUROSPEECH03*, pages 381–384, Geneve, Switzerland.

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

Ding Liu and Daniel Gildea. 2007. Source-language features and maximum correlation training for machine translation evaluation. In *Proc. of the NAACL-HLT*, pages 41–48, Rochester NY, USA.

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

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

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

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

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

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

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

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

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