Is this translation error critical?: Classification-based Human and Automatic Machine Translation Evaluation Focusing on Critical Errors

Katsuhito Sudoh*, **Kosuke Takahashi, Satoshi Nakamura**
Nara Institute of Science and Technology (NAIST)
8916-5 Takayamacho, Ikoma, Nara 630-0192, Japan
*PRESTO, Japan Science and Technology Agency (JST)
{sudoh,kosuke.takahashi.th0,s-nakamura}@is.naist.jp

Abstract

This paper discusses a classification-based approach to machine translation evaluation, as opposed to a common regression-based approach in the WMT Metrics task. Recent machine translation usually works well but sometimes makes critical errors due to just a few wrong word choices. Our classification-based approach focuses on such errors using several error type labels, for practical machine translation evaluation in an age of neural machine translation. We have made additional annotations on the WMT 2015-2017 Metrics datasets with fluency and adequacy labels to distinguish different types of translation errors from syntactic and semantic viewpoints. We present our human evaluation criteria for the corpus development and automatic evaluation experiments using the corpus. The human evaluation corpus will be publicly available at https://github.com/ksudoh/wmt15-17-humaneval.

1 Introduction

Most machine translation (MT) studies still evaluate their results using BLEU (Papineni et al., 2002) because of its simple, language-agnostic, and model-free methodology. Recent remarkable advances in neural MT (NMT) have cast an important challenge in its evaluation; NMT usually generates a fluent translation that cannot always be evaluated precisely by simple surface-based evaluation metrics like BLEU.

A recent trend in the MT evaluation is to use a large-scale pre-trained model like BERT (Devlin et al., 2019). Shimanaka et al. (2019) proposed BERT Regressor based on sentence-level regression using a fine-tuned BERT model, as an extension of their prior study using sentence embeddings (Shimanaka et al., 2018). Zhang et al. (2020) proposed BERTScore based on hard token-level alignment using cosine similarity of contextualized token embeddings. Zhao et al. (2019) proposed MoverScore based on soft token-level alignment using Word Mover's Distance (Kusner et al., 2015). Sellam et al. (2020) proposed BLEURT that incorporates auxiliary task signals into the pre-training of a BERT-based sentence-level regression model. These methods aim to evaluate a translation hypothesis using the corresponding reference with a high correlation to human judgment.

The evaluation of this kind of MT evaluation, often called *meta-evaluation*, is usually based on some benchmarks. The meta-evaluation in the recent studies uses the WMT Metrics task dataset consisting of human judgment on MT results. The human judgment is given in the form of Human Direct Assessment (DA) (Graham et al., 2016), a 100-point rating scale. The Human DA results are standardized into *z-scores* (human DA scores, hereinafter) and used as the evaluation and optimization objective of regression-based MT evaluation methods. Recent MT evaluation methods achieved more than 0.8 in Pearson correlation on WMT 2017 test set[1]. However, Takahashi et al. (2020) reported a weaker correlation in low human DA score ranges. Such a finding suggests the difficulty of the MT evaluation on low-quality results.

In this work, we focus on the problem in the evaluation of low-quality translations that cause serious misunderstanding. Judging erroneous translations in the 100-point rating scale would be very difficult and unstable, because the extent of errors cannot be mapped easily into a one-dimensional space. Suppose we are evaluating a translation hypothesis, (1) *It is our duty to remain at his sides* with its reference, *It is not our duty to remain at his sides*.[2] The

[1] The correlation got worse in the newer WMT datasets (Ma et al., 2018, 2019) due to noise in human judgement (Sellam et al., 2020).

[2] This example is taken from the Metrics dataset of WMT

difference in this example is just in one missing word *not* in the hypothesis, but it may cause a serious misunderstanding. Such translation errors are considered as critical ones by professional translators. There are several metrics for translation quality assessment (QA) proposed in the translators' community, such as LISA QA Metric[3] and Multidimentional Quality Metrics (MQM)[4]. These metrics use a couple of error seriousness categories (Minor, Major, Critical) in several viewpoints, such as mistranslation, accuracy, and terminology. The missing negation is a kind of critical error. Nevertheless, most existing automatic MT evaluation metrics fail to penalize such errors. Human DA is also difficult from this viewpoint. Suppose we have other translation hypotheses, (2) *He bought some bags at a duty-free store.* and (3) *Not is to duty remain it sides his at.* for the same reference. We can easily identify these hypotheses are wrong. However, evaluating them together with (1) in the same 100-point rating scale by mapping these differences into one dimension is not trivial.

This work pursues a classification-based human and automatic MT evaluation based on the multidimensional evaluation. Current NMT technologies would still be far from the level of professional human translators but are also utilized in various applications. MT in practical applications should be evaluated as same as human translations by practical metrics, not just by incremental and engineering-oriented metrics like BLEU.

We propose a classification-based MT evaluation framework motivated by the discussion about critical errors. In human evaluation, we use conventional evaluation dimensions of fluency and adequacy (LDC, 2005) and define several categories different from a conventional 1-5 Likert scale. We developed a corpus with such additional annotations on WMT Metrics dataset and found that human DA scores penalize incomprehensible and unrelated MT hypotheses more than those with other critical errors that cause serious misunderstanding and contradiction. We then implemented a classification-based automatic MT evaluation using the corpus and conducted experiments on the WMT Metrics test set.

2 Related Work

MT evaluation has evolved along with the advance of MT technologies. White et al. (1994) reviewed some attempts of human evaluation and presented adequacy, fluency, and comprehension results in the early 1990s. The Quality Panel approach presented in their paper was motivated by the evaluation of human translations, but it was finally abandoned due to human evaluation difficulties. Callison-Burch et al. (2007) presented meta-evaluation of the MT evaluation in WMT shared tasks. According to the findings there, the WMT shared tasks had employed ranking-based human evaluation for a while. Snover et al. (2006) defined Human-targeted Translation Edit Rate (HTER) that measures the translation quality by the required number of post-edits on a translation hypothesis. Denkowski and Lavie (2010) and Graham et al. (2012) discussed the differences among those human evaluation approaches. Graham et al. (2016) proposed human DA for the MT evaluation, and DA has been used as standard human evaluation in recent WMT Metrics tasks.

There is another line of human MT evaluation studies focusing on semantics. Lo and Wu (2011) proposed MEANT and its human evaluation variant HMEANT based on semantic frames. Birch et al. (2016) proposed HUME based on a semantic representation called UCCA. This kind of fine-grained semantic evaluation requires some linguistic knowledge for annotators but enables explainable evaluation instead. However, the meaning of the sentence can be changed by small changes, as discussed later in section 3. Looking at sub-structures and using their coverage in the MT evaluation may suffer from this problem.

One recent approach has been proposed by Popoviĉ (Popovic, 2020; Popović, 2020). Her work analyzed the differences between comprehensibility and adequacy in machine translation outputs. The human annotations in her work are major and minor errors in comprehensibility and adequacy on words and phrases. These fine-grained annotations are helpful for detailed translation error detection. The focus of our work is different; we are going to develop sentence-level MT evaluation through simpler human and automatic evaluation schemes.

In this work, we suggest revisiting the classification-based evaluation with fluency and

2015.

[3] http://producthelp.sdl.com/SDL_TMS_2011/en/Creating_and_Maintaining_Organizations/Managing_QA_Models/LISA_QA_Model.htm

[4] https://www.dfki.de/en/web/research/projects-and-publications/publications-overview/publication/7717/

adequacy, for *absolute* human and automatic evaluation. DA-based human evaluation is beneficial in demonstrating the correlation with automatic evaluation metrics. However, it is not very intuitive in the evaluation of different kinds of translation errors.

Our work is also related to some studies using semantic equivalence and contradiction. BLEURT (Sellam et al., 2020) employed NLI in its pre-training phase. NLI includes contradiction identification, which should also contribute to the MT evaluation. BLEURT has revealed its advantage in the example shown in Table 1. Kryściński et al. (2019) proposed a weakly-supervised method for training an abstractive summarization model using adversarial summaries to improve the factual consistency between a source document and a summary. They also focused on an NLI-like semantic classification for their adversarial training. Classification-based automatic MT evaluation models can be trained similarly, using related and adversarial data.

3 Critical Translation Errors

The main focus of this work is to penalize critical errors in translation hypotheses that cause serious misunderstanding. This kind of translation errors must be avoided, as well as possible.

Suppose we have some translation hypotheses with their reference, *The Pleiades cluster is situated 445 light-years from Earth*[5]. The translation hypotheses are artificial ones with some adversarial edits over the reference, as shown in the second column of Table 1. The hypothesis hyp1 is a paraphrase, hyp2 and hyp3 have errors on "light-years", hyp4 has a wrong negation, hyp5 to hyp7 have errors on named entities, hyp8 is a shuffled word sentence, and hyp9 would come from a completely different sentence; the hypotheses have non-trivial problems except hyp1.

We put automatic evaluation scores in the table using BLEU-4[6], chrF[7], BERTScore[8], and BLEURT[9]. hyp9 is correctly penalized by all the metrics, but the other results are mixed. BLEU-4 penalizes hyp1 and hyp3 more than the others. chrF and BERTScore penalize hyp3. BLEURT penalizes hyp4 and gives lower scores on hyp2 and hyp5-7 than BERTScore. BLEU-4, BERTScore, and BLEURT penalize hyp8, while chrF gives the same score on it as hyp3. Here, we would regard hyp4, hyp8, and hyp9 as bad translations. However, we cannot identify the other erroneous translation just using the automatic scores. These observations suggest that current evaluation metrics do not always capture these critical translation errors by one or two wrong word choices. Recent NMT sometimes generates translations competitive with human translators, so they should be evaluated as same as human translations in practice.

On the other hand, MT sometimes generates incomprehensible sentences with various kind of errors, even though NMT works much better than conventional statistical MT, especially in fluency. Such incomprehensible translations are also very problematic as well as content errors in easy-to-understand and fluent translations.

However, it is not easy to penalize both of them in a single evaluation criterion. Existing automatic evaluation methods often fail to penalize content errors, although they work well for incomprehensible and unrelated sentences, as revealed by the adversarial examples in Table 1. In this work, we aim to differentiate these errors motivated by the conventional evaluation dimensions of fluency and adequacy (LDC, 2005).

4 Human Evaluation Corpus

We have developed a new human evaluation corpus from the viewpoints of fluency and adequacy. The evaluation corpus is available at GitHub[10] under CC BY-NC-SA 4.0[11]. In this section, we present the details of the corpus. Here, the human evaluation is designed in *monolingual* way; an MT hypothesis is evaluated against only its reference, supposing the reference is semantically equivalent to the source language input.

We made a contract with a linguistic data development company to conduct the human evaluation[12] with three annotators who are native speaker

[5] This example is taken from the Metrics dataset of WMT 2017.

[6] sacrebleu fingerprint: BLEU+case.lc+numrefs.1+smooth.exp+tok.13a+version.1.4.8

[7] sacrebleu fingerprint: chrF2+case.lc+numchars.6+numrefs.1+space.False+version.1.4.8

[8] Authors' implementation https://github.com/Tiiiger/bert_score with fingerprint: roberta-large_L17_no-idf_version=0.3.2(hug_trans=2.8.0)-rescaled

[9] Authors' implementation https://github.com/google-research/bleurt

[10] https://github.com/ksudoh/wmt15-17-humaneval

[11] https://creativecommons.org/licenses/by-nc-sa/4.0/

[12] The human evaluation was conducted without formal ethical review.

ID	Hypothesis	BLEU-4	chrF	BERTScore	BLEURT
ref	The Pleiades cluster is situated 445 light-years from Earth.	1.0	1.0	1.000	0.940
hyp1	The Pleiades cluster is ~~situated~~ 445 light-years far from Earth.	0.423	0.8	0.932	0.800
hyp2	The Pleiades cluster is situated 445 years from Earth.	0.658	0.8	0.854	0.670
hyp3	The Pleiades cluster is ~~situated~~ 445 light from the Earth.	0.336	0.6	0.617	0.698
hyp4	The Pleiades cluster is not situated 445 light-years from Earth.	0.702	0.9	0.892	0.028
hyp5	The Pleiades cluster is situated 345 light-years from Earth.	0.658	0.9	0.946	0.709
hyp6	The Pleiades cluster is situated 445 light-years from Mars.	0.783	0.9	0.909	0.640
hyp7	The Hyades cluster is situated 445 light-years from Earth.	0.783	0.9	0.891	0.556
hyp8	Is Earth from Pleiades the light-years situated cluster 445.	0.071	0.6	0.393	-0.659
hyp9	Turn off the light for saving the Earth.	0.085	0.2	0.039	-1.55

Table 1: Examples of automatic MT evaluation on adversarial examples. Underlines and strikethroughs represent differences from the reference.

of English and had work experiences of translation into English. We provide a set of English sentence pairs to the annotators: translation hypotheses and the corresponding references. No specific training was conducted before the evaluation. The annotators can ask questions to a moderator in the company, and the moderator asked them to the first author. The annotators conducted the evaluation independently, referring to the evaluation criteria below.

4.1 Dataset

We chose the WMT 2015-2017 Metrics datasets to give additional annotations. The MT results in the dataset and the corresponding human DA scores have been used in many existing automatic MT evaluation studies. The total number of pairs of hypothesis and reference sentences was 9,280, consisting of 2,000 pairs from WMT 2015, 3,360 pairs from WMT 2016, and 3,920 pairs from WMT 2017 datasets.

4.2 Evaluation Criteria

We propose the following evaluation criteria in fluency and adequacy, shown in Tables 2 and 3, respectively.

4.2.1 Fluency

The fluency criteria in Table 2 extend conventional ones by LDC (2005), with a *comprehension* viewpoint in the lower range. The lowest judgment *Incomprehensible* corresponds to LDC's fluency criterion "1: Incomprehensible," but is not limited to disfluency problems. The category *Poor* means the difficulty of comprehension. The other categories are defined mainly from a fluency viewpoint.

When a sentence is incomprehensible such as hyp8 in Table 1, we cannot evaluate its contents in the adequacy evaluation. On the other hand, hyp9 is not related to the reference and should be judged as a critical error in adequacy, even though it is easy-to-understand and looks fluent. These criteria were also motivated by the *acceptability* criteria (Goto et al., 2011). By the acceptability criteria, a hypothesis that lacks important information (i.e., its adequacy is not 5 in the five-point scale) is always judged as the worst, and better labels are given according to grammatical correctness and fluency.

4.2.2 Adequacy

Our adequacy criteria in Table 3 are different from the conventional ones (LDC, 2005) that focused on the amount of important information. We defined the adequacy of a translation hypothesis focusing

Category	Explanation
Incomprehensible (F)	The sentence is not comprehensible.
Poor (D)	Some contents are not easy to understand by typographical/grammatical errors and problematic expressions
Fair (B)	All the contents are easy to understand in spite of some typographical/grammatical errors
Good (A)	All the contents are easy to understand and free from grammatical errors, but some expressions are not very fluent
Excellent (S)	All the contents are easy to understand, and all the expressions are flawless

Table 2: Evaluation criteria in *Fluency*. Labels in parentheses are the ones used in the evaluation corpus.

on the delivery of the correct information, based on the discussion in section 3. Our criteria put more focus on possible *misunderstanding* by a translation hypothesis; we consider a translation may cause serious misunderstanding even if most parts of the translations are correct.

First, we use the category *Incomprehensible* for such hypotheses that are also classified into *Incomprehensible* in fluency. Then, we divide critical content errors into three types: *Unrelated*, *Contradiction*, and *Serious*. *Unrelated* indicates the unrelatedness, as shown by hyp9 in Table 1. It is expected to appear in poor translations in a very low-resourced condition. The category *Contradiction* indicates the contradiction with the reference, such as a negation flip at hyp4 and a number error at hyp5 in Table 1. This label was motivated by the task of natural language inference (NLI), which has also been used for the pre-training of MT evaluation (Sellam et al., 2020). The category *Serious* covers the other kind of serious content errors such as hyp6, and hyp7 in Table 1. These hypotheses deliver somewhat related but different information compared to the reference. The intermediate categories *Fair* and *Good* are used for major and minor errors, respectively.

Category	Explanation
Incomprehensible (N)	The contents cannot be understood due to fluency and comprehension issues, so the hypothesis is not eligible for the adequacy evaluation.
Unrelated (O)	The hypothesis delivers information that is *not related* to the reference
Contradiction (C)	The hypothesis delivers information that *contradicts* the reference
Serious (F)	The hypothesis delivers information that may cause serious misunderstanding due to some content errors but does not contradict the reference
Fair (B)	The hypothesis has some problems in its contents but does not cause a serious misunderstanding
Good (A)	The hypothesis has some minor problems in its contents that do not make a misunderstanding
Excellent (S)	The hypothesis delivers information equivalent to the reference.

Table 3: Evaluation criteria in *Adequacy*. Labels in parentheses are the ones used in the evaluation corpus.

4.3 Analyses

We conducted some analyses on the human evaluation corpus mainly in the differences among the three annotators.

4.3.1 Annotation Bias

We analyzed annotation differences among the three annotators (named A, B, and C), especially their labeling biases. Tables 4 and 5 show the annotation distributions for the three annotators on fluency and adequacy, respectively. We can see some differences among the annotators; for example, annotator B was very strict for using the best category *Excellent* in both dimensions, and annotator C gave more bad labels (*Contradiction* and *Serious*) than the others.

On average, the translation hypotheses in the WMT Metrics dataset for 2015-2017 still include many translation errors. The error tendency would be different on newer data consisting of many recent neural MT results. It is worth investigating recent MT results in future studies.

Fluency	A	B	C	Ave.
Incomprehensible	0.098	0.099	0.111	0.103
Poor	0.167	0.220	0.181	0.189
Fair	0.356	0.406	0.222	0.328
Good	0.124	0.240	0.219	0.195
Excellent	0.254	0.035	0.266	0.185

Table 4: Annotation distributions for the three annotators (fluency).

Adequacy	A	B	C	Ave.
Incomprehensible	0.098	0.099	0.098	0.098
Unrelated	0.004	0.001	0.011	0.005
Contradiction	0.009	0.019	0.086	0.038
Serious	0.205	0.187	0.311	0.234
Fair	0.374	0.343	0.146	0.288
Good	0.233	0.296	0.271	0.267
Excellent	0.076	0.005	0.076	0.069

Table 5: Annotation distributions for the three annotators (adequacy).

4.3.2 Comparison with Human Direct Assessment Scores

We compared our human evaluation labels with the human DA scores (standardized z-scores) given in the WMT Metrics data. Tables 6 and 7 show the mean and standard deviation values of human DA scores for each human evaluation label.

The human DA score ranges of the fluency and adequacy labels had almost the same partial orders among different annotators, although they still reflect the annotation bias shown in Tables 4 and 5; annotator B had a higher standard in fluency evaluation than the others.

One important finding here is the differences among the adequacy categories *Incomprehensible*, *Unrelated*, *Contradiction* and *Serious* in Table 7. The sentences with *Unrelated* were scored the worst by the human DA. However, critical content errors suggested by the labels *Contradiction* and *Serious* were penalized less than the ones with *Incomprehensible* and *Unrelated*. Such content errors should also be identified as critical translation errors in practice.

4.3.3 Inter-annotator Agreement

We also measured pairwise agreement among the three annotators using the κ coefficient (Carletta, 1996) and label concordance rate. The results are shown in Table 8. The inter-annotator agreement was not high enough but κ values are also comparable to the previous studies on older WMT datasets (Callison-Burch et al., 2007; Denkowski and Lavie, 2010)[13]. The agreement in fluency was lower than that in adequacy, especially on A-B and B-C, due to very high fluency standard of the annotator B. The agreement would improve with careful pre-annotation training and more example-based evaluation guidelines, because the annotators gave us feedback about the difficulty in discrimination among different categories.

5 Experiments

We conducted experiments using the evaluation corpus, to investigate the performance of automatic classification-based MT evaluation.

5.1 Experimental Setup

5.1.1 Data

Among the evaluation corpus, we reserved the WMT 2017 portion (3,920 samples; 560 for each language pair — cs-en, de-en, fi-en, lv-en, ru-en, tr-en, and zh-en) for the test set, chose 536 samples randomly for the development set, and used the remained 4,824 samples for the training set.

We took agreements among the three different annotators for the experiments by the following heuristics.

- If two or three annotators gave the same label, it was used as the agreement.
- If the annotators' judgment were different from each other, the worst label was used as the agreement. The label order was *Incomprehensible* $<$ *Poor* $<$ *Fair* $<$ *Good* $<$ *Excellent* for fluency and *Contradiction* $<$ *Serious* $<$ *Incomprehensible* $<$ *Unrelated* $<$ *Fair* $<$ *Good* $<$ *Excellent* for adequacy[14].

Tables 9 and 10 show the label statistics on the training, development, and test sets after applying the heuristics.

5.1.2 Automatic Evaluation Method

We used a simple sentence-level automatic MT evaluation framework, which takes hypothesis and reference sentences as the input and predicts the label. Since the task in the experiments was classification, the evaluation model was trained with

[13] Note that we had three annotators who evaluated all the sentences.

[14] We used this heuristic order because of the importance of content errors suggested by *Contradiction* and *Serious*.

Fluency	A	B	C
Incomprehensible	-0.644 (0.371)	-0.692 (0.356)	-0.649 (0.378)
Poor	-0.421 (0.408)	-0.420 (0.399)	-0.400 (0.418)
Fair	-0.079 (0.478)	0.019 (0.474)	-0.129 (0.449)
Good	0.165 (0.479)	0.408 (0.485)	0.122 (0.467)
Excellent	0.428 (0.524)	0.644 (0.465)	0.427 (0.521)

Table 6: Mean (standard deviation) of Direct Assessment scores for labels by the three annotators (fluency)

Adequacy	A	B	C
Incomprehensible	-0.646 (0.369)	-0.692 (0.356)	-0.662 (0.373)
Unrelated	-0.990 (0.367)	-0.926 (0.415)	-0.963 (0.363)
Contradiction	-0.370 (0.460)	-0.366 (0.468)	-0.200 (0.501)
Serious	-0.453 (0.438)	-0.499 (0.425)	-0.279 (0.473)
Fair	-0.076 (0.435)	-0.092 (0.417)	-0.029 (0.414)
Good	0.417 (0.361)	0.414 (0.363)	0.347 (0.414)
Excellent	0.814 (0.278)	0.839 (0.294)	0.756 (0.327)

Table 7: Mean (standard deviation) of Direct Assessment scores for labels by the three annotators (adequacy)

Metric		A-B	A-C	B-C
Fluency	κ	0.2860	0.3773	0.2489
	r	0.4512	0.5113	0.4014
Adequacy	κ	0.3947	0.2684	0.2774
	r	0.5459	0.5870	0.5752

Table 8: Inter-annotator agreement in κ coefficient and label concordance rate (r) on our human evaluation corpus. The fluency metric has five categories and the adequacy metric has seven categories.

Fluency	Training	Dev.	Test
Incomprehensible	545	74	282
Poor	992	96	602
Fair	1,655	196	1,341
Good	808	80	899
Excellent	824	90	796

Table 9: Label statistics of *fluency* dataset.

Adequacy	Training	Dev.	Test
Incomprehensible	617	87	350
Unrelated	15	2	19
Contradiction	93	6	40
Serious	1,161	108	717
Fair	1,433	162	1,441
Good	1,208	143	1,165
Excellent	297	28	188

Table 10: Label statistics of *adequacy* dataset.

the classification objective, softmax cross-entropy over the category distribution. We trained and used independent models for fluency and adequacy.

We implemented the evaluator using HuggingFace Transformers[15] and its pre-trained RoBERTa model (`roberta-large`) (Liu et al., 2019). The model was fine-tuned to predict a label through an additional feed-forward layer taking the vector for `[CLS]` token as the input, using a softmax cross-entropy loss. Due to the label imbalance shown in Tables 9 and 10, we applied a sample-wise loss scaling with weights that were inversely proportional to the number of training instances with the labels. A label weight for a category c was defined as:

$$w_c = \sqrt{\frac{\max_{c' \in \mathcal{C}} \mathrm{count}_{c'}}{\mathrm{count}_c}}, \tag{1}$$

where $\mathcal{C}$ is a set of categories.

We employed the Adam optimizer (Kingma and Ba, 2015) and continued the training for 30 epochs with the initial learning rate of 1e-5. We tried different minibatch sizes (4, 8, 16) and dropout rates in the additional feed-forward layer (0.1, 0.3, 0.5, 0.75)[16], and used the ones resulting in the best classification accuracy on the development set: 4 and 0.75 for fluency, 8 and 0.5 for adequacy, respectively.

[15] `https://github.com/huggingface/transformers`

[16] The dropout rate in RoBERTa was kept unchanged from its default value of 0.1. We also tried to increase it in the pilot test, but that resulted worse.

ref\pred	Inc.	Poor	Fair	Good	Exc.
Incomprehensible	**206**	45	22	8	1
Poor	45	**266**	250	43	4
Fair	15	134	**782**	358	52
Good	2	11	187	**560**	139
Excellent	0	2	35	306	**453**

Table 11: Confusion matrix in *fluency* prediction. The **bold** numbers represent correct predictions. The overall classification accuracy was 0.578.

Fluency	Precision	Recall	F1-score
Incomprehensible	0.769	0.730	0.749
Poor	0.581	0.438	0.499
Fair	0.613	0.583	0.598
Good	0.439	0.623	0.515
Excellent	0.698	0.569	0.627
Ave.	0.620	0.589	0.598

Table 12: Precision, recall, and F1-score in *fluency* prediction.

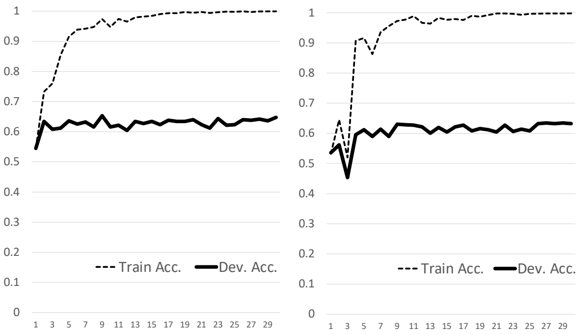
5.2 Results

We show the statistics of the prediction results by a confusion matrices and precision/recall/F1-scores. Tables 11 and 12 are from the fluency prediction, and Tables 13 and 14 are from the adequacy prediction.

In the fluency prediction, the classification accuracy on the test set was 0.578 (2,267 correct predictions out of 3,920), and that on the training and development sets was 0.999 and 0.647, respectively. Most of the incorrect predictions were in adjacent categories, and the fraction of serious misrecognition in distant categories (*Incomprehensible* → {*Good*, *Fair*}, *Poor* → *Excellent*, *Good* → *Incomprehensible*, and *Excellent* → {*Incomprehensible*, *Poor*}) was not so large (0.43%; 17 out of 3,920).

The prediction performance in Table 12 suggests the best and worst categories (*Excellent* and *Incomprehensible*) can be predicted more accurately than the intermediate categories.

In the adequacy prediction, the classification accuracy on the test set was 0.600 (2,351 correct predictions out of 3,920), and the results on the training and development sets were 0.998 and 0.632, respectively. The prediction of less frequent categories (*Unrelated* and *Contradiction*) did not work well despite the instance weighting in training. The result suggests we should use more negative examples in training for more accurate predictions on them. The prediction performance in Table 14 suggests the hypotheses with *Incomprehensible* can be identified more accurately than the others. Predictions of the other categories were still difficult. However, 93.5% of the hypotheses with the predicted label *Excellent* were good translations labeled *Excellent* or *Good* (144 out of 154); this finding would be beneficial in practice. The most serious confusion in this result was between *Serious* (critical) and *Fair* (okay). More fine-grained discrimination is needed to judge them.



(a) Fluency (b) Adequacy

Figure 1: Learning curves in classification accuracy over training epochs.

Figure 1 (a) and (b) show the learning curves. The training set accuracy was almost saturated around 20 training epochs, but the development set accuracy was not stable until 30 epochs.

In summary, these experiments suggest our classification-based MT evaluation with absolute categories is promising, while we still need more *negative* examples. More data collections, including data augmentation, would be helpful, along with a further investigation of prediction models.

6 Conclusions

In this paper, we present our approach to classification-based human and automatic MT evaluation, focusing on critical translation errors in MT outputs. We revisited the use of fluency and adequacy metrics with some modifications on evaluation criteria, motivated by our thoughts on the critical content errors.

We developed a human evaluation corpus based on the criteria using the WMT Metrics dataset, which will be publicly available upon publication. Our corpus analyses revealed the human DA penalizes unrelated and incomprehensible hypotheses much more than contradiction and other critical errors in the content. We also conducted automatic

r\p	Inc.	Unr.	Con.	Ser.	Fair	Good	Exc.
Incomprehensible	**224**	0	0	83	38	4	1
Unrelated	0	**1**	0	13	5	0	0
Contradiction	0	0	**8**	9	13	10	0
Serious	37	0	8	**385**	242	45	0
Fair	29	0	13	237	**878**	274	10
Good	4	0	9	20	302	**771**	59
Excellent	0	0	0	1	6	97	**84**

Table 13: Confusion matrix in *adequacy* prediction. The **bold** numbers represent correct predictions. The overall classification accuracy was 0.600.

Adequacy	Precision	Recall	F1-score
Incomprehensible	0.762	0.640	0.696
Unrelated	1.000	0.053	0.100
Contradiction	0.211	0.200	0.205
Serious	0.515	0.537	0.526
Fair	0.592	0.609	0.600
Good	0.642	0.662	0.652
Excellent	0.545	0.447	0.491
Ave.	0.609	0.450	0.467

Table 14: Precision, recall, and F1-score in *adequacy* prediction.

MT evaluation experiments using the human evaluation corpus and achieved around 60% classification accuracy both in fluency and adequacy.

Our future work includes further development of human evaluation corpora that are not limited to WMT Metrics data, and data augmentation methods to tackle the label imbalance problem. It is also promising to apply the classification-based automatic MT evaluation to the neural MT training.

Acknowledgments

The authors would like to thank anonymous reviewers for their comments and suggestions. This work is supported by JST PRESTO (JPMJPR1856).

References

Alexandra Birch, Omri Abend, Ondřej Bojar, and Barry Haddow. 2016. HUME: Human UCCA-based evaluation of machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1274, Austin, Texas. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Michael Denkowski and Alon Lavie. 2010. Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks . In *Proceedings of the Ninth Biennial Conference of AMTA 2010.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin Tsou. 2011. Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 559–578.

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

Yvette Graham, Timothy Baldwin, Aaron Harwood, Alistair Moffat, and Justin Zobel. 2012. Measurement of progress in machine translation. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 70–78, Dunedin, New Zealand.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the Third International Conference on Learning Representations*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the Factual Consistency of Abstractive Text Summarization. *arXiv preprint arXiv: 1910:12840*.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From Word Embeddings To Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

LDC. 2005. Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations Revision 1.5, January 25, 2005 . Technical report, Linguistic Data Consortium.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv: 1907:11692*.

Chi-kiu Lo and Dekai Wu. 2011. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 220–229, Portland, Oregon, USA. Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maja Popovic. 2020. On the differences between human translations. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 365–374, Lisboa, Portugal. European Association for Machine Translation.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2019. Machine Translation Evaluation with BERT Regressor. *arXiv preprint arXiv: 1907.12679*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea, Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-2006)*.

Kosuke Takahashi, Katsuhito Sudoh, and Satoshi Nakamura. 2020. Automatic machine translation evaluation using source language inputs and cross-lingual language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3553–3558, Online. Association for Computational Linguistics.

John S. White, Theresa A. O'Connell, and Francis E. O'Mara. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the Eighth International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.