Japanese to English/Chinese/Korean Datasets for Translation Quality Estimation and Automatic Post-Editing

Atsushi Fujita and **Eiichiro Sumita**

National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
firstname.lastname@nict.go.jp

Abstract

Aiming at facilitating the research on quality estimation (QE) and automatic post-editing (APE) of machine translation (MT) outputs, especially for those among Asian languages, we have created new datasets for Japanese to English, Chinese, and Korean translations. As the source text, actual utterances in Japanese were extracted from the log data of our speech translation service. MT outputs were then given by phrase-based statistical MT systems. Finally, human evaluators were employed to grade the quality of MT outputs and to post-edit them. This paper describes the characteristics of the created datasets and reports on our benchmarking experiments on word-level QE, sentence-level QE, and APE conducted using the created datasets.

1 Introduction

Technologies of machine translation (MT) have been dramatically improved in the last decades; however, the strict requirements for high-quality translations in real-world applications (Hutchins and Somers, 1992) have not yet fulfilled by MT systems alone.[1] Thus, in practice, techniques of computer-aided translation (CAT) have been widely used to provide satisfiable translations for such requirements. For instance, manual post-editing of MT outputs has become a prevalent translation work-flow in translation services (ISO/TC27, 2017). Quality estimation (QE) of MT outputs also plays a critical role in CAT to reduce human effort, thereby increasing productivity (Specia et al., 2010).

To facilitate and encourage the research on QE tasks concerning several different levels of granularity, i.e., word, phrase, sentence, and document levels, and automatic post-editing (APE), WMT workshops and conferences (henceforth, WMT) have created datasets specialized for these tasks (Bojar et al., 2014, 2015, 2016, 2017), mainly focusing on European languages.[2] As a result, they have successfully led to the rapid enhancement of QE/APE technologies.

However, to the best of our knowledge, such a resource for Asian languages have never emerged, and QE/APE for Asian languages have been less studied. Aiming at facilitating this line of research, we have created new datasets[3] consisting of the 5-tuples shown in Figure 1. While the tuples of first two elements, i.e., source text and human translation, compose ordinary parallel corpus used to train (data-driven) MT systems, the remaining three are specific to this kind of QE/APE datasets. So far, we have regarded Japanese (Ja) as the source language, and English (En), Chinese (Zh), and Korean (Ko) as the target languages. In addition to cover these new language pairs, we also aim to improve our speech translation service[4] with QE/APE technologies. To this end, we have used actual utterances for the source texts, accumulated by the speech translation service, with our best effort to clean and anonymize the data.

In the remainder of this paper, we first describe the procedure of creating our QE/APE datasets for Ja→En, Ja→Zh, and Ja→Ko translation tasks in Section 2. Then, in Section 3, we present statistics of the created datasets, observations, and remaining issues. Section 4 describes our benchmarking

[1] Bar-Hillel (1951) even mentioned that the fully automatic high-quality translation is not only unrealistic, but also theoretically impossible.

[2] Only the exception is Chinese-to-English in 2017 (Bojar et al., 2017).

[3] NICT QE/APE Dataset, http://att-astrec.nict.go.jp/en/product/

[4] VoiceTra, http://voicetra.nict.go.jp/en/

Proceedings of the 4th Workshop on Asian Translation, pages 79–88,
Taipei, Taiwan, November 27, 2017.

Component	Example
src: Source segment in Japanese	片道だけで買えますか。
ref: Human translation	May I get it for one way?
hyp: MT output	Can I buy just one way?
grade: Quality grade of MT output	B (∈ {S, A, B, C, D})
pe: Manually post-edited MT output	Can I just buy a one way ticket?

Figure 1: Example record in our QE/APE datasets (see Section 2.4 for the definition of *grade*).

experiments on word-level QE, sentence-level QE, and APE conducted using the created datasets. Finally, Section 5 summarizes this paper.

2 Procedure of Corpus Construction

We have created our QE/APE datasets, regarding Japanese as the source language. We have so far regarded English, Chinese, and Korean as the target languages, considering that the speakers of these languages hold the largest proportion of visitors to Japan (Japan National Tourism Organization, 2017). Following the procedure in previous studies (Snover et al., 2006; Potet et al., 2012) and practices in WMT (Bojar et al., 2014, 2015, 2016), we determined the following five-step process.

1. Collecting Japanese utterances (*src*)
2. Generation of MT outputs (*hyp*)
3. Manual translation (*ref*)
4. Manual grading of MT output (*grade*)
5. Manual post-editing of MT output (*pe*)

For the latter three tasks (detailed in Sections 2.3, 2.4 and 2.5, respectively), we allocated adult native speakers of the target language who also understand Japanese.

2.1 Collecting Japanese utterances (*src*)

First, we collected the following two sets of utterances in Japanese that have been used with our speech translation service.

Travel-related utterances (*travel*): From the log data that our speech translation service accumulates, we randomly sampled 20,000 identical transcribed segments[5] that were identified as Japanese by its automatic speech recognition (ASR) module. Most segments were spoken language and related to travel and tourism, even though we had no restriction to the input of our users.

[5]In this paper, we refer to each utterance as "segment," as one utterance may contain more than one sentence.

Utterances in hospital (*hospital*): We employed the role-play dialogs of health care providers, such as doctors and nurses, and patients, containing 2,225 identical segments of utterances. They were surely spoken language, although they were manually written and more formal than those in the *travel* domain.

We have been examining the installation of our speech translation service into several practical situations where such system helps cross-lingual communication between humans. For this purpose, we have manually created role-play dialogs between Japanese and non-Japanese speakers. The *hospital* data is one of them.

The extracted segments, especially those in the *travel* domain, include ungrammatical ones, non-understandable ones, and those containing inappropriate expressions with respect to social standards. We therefore asked a native Japanese speaker to filter out such segments. As a result, 8,783 and 1,676 segments in the *travel* and *hospital* domains were retained, respectively.

Many segments do not have an explicit subject, as Japanese is a pro-drop language; even obligatory arguments can be missing. For instance, in the *src* segment in Figure 1, both the subject "I" and the direct object "ticket" are omitted. However, we cannot recover them as our speech translation service does not record any discourse elements of individual utterances.

2.2 Generation of MT outputs (*hyp*)

The collected Japanese segments (*src*) were then translated by our in-house MT systems, which implement a phrase-based statistical MT (Koehn, 2009). The Ja→En translations were obtained in 2013, with the system trained on 736k sentence pairs. The Ja→Zh and Ja→Ko translations were generated later in 2016, with the systems trained on 1.44M and 1.40M sentence pairs, respectively.

Table 1 summarizes the statistics of *src* and *hyp*. These segments are relatively shorter than sentences in written texts, such as news articles and patent documents.

Table 1: Statistics of the Japanese *src* and *hyp* in each target language.

Partition	Unit	*travel* (8,783 segments)				*hospital* (1,676 segments)			
		Total	Min	Avg.	Max	Total	Min	Avg.	Max
Japanese *src*	character	105,606	2	12.0	49	33,979	5	20.3	71
English *hyp*	word	44,604	1	5.1	28	14,844	1	8.9	29
Chinese *hyp*	character	65,710	2	7.5	30	21,974	3	13.1	41
Korean *hyp*	character	94,578	2	10.8	48	30,283	3	18.1	60

Table 2: Grading criterion for human evaluators.

Grade	Summary	Description
S	Perfect	Information of the source text has been completely translated. There are no grammatical errors in the target text. Lexical choice and phrasing are natural even from a native speaker point-of-view.
A	Good	The information of the source text has been completely translated and there are no grammatical errors in the target text, but lexical choice and phrasing are slightly unnatural.
B	Fair	There are some minor errors in the target text of less important textual information, but the meaning of the source text can be easily understood.
C	Acceptable	Important parts of the source text are omitted or could not be translated correctly, but the meaning of the source text can still be understood with some efforts.
D	Incorrect	The meaning of the source text is incomprehensible from target text.

2.3 Manual translation (*ref*)

Reference translations were manually given, referring only to the source segments (*src*). As each *src* was not attributed with its specific context, we asked the translators to imagine some context as long as it is reasonable considering the domain. On the contrary, we also asked to avoid adding too much contents that cannot be specified only from the *src*. For the *src* which has more than one interpretation, only one translation is given rather than enumerating all the possible interpretations.

2.4 Manual grading of MT output (*grade*)

The quality of MT output (*hyp*) with respect to its source (*src*) was graded according to a standard presented in Table 2, which is compatible[6] with the "Acceptability" criterion in Goto et al. (2013). In case the evaluator cannot understand the meaning of *src*, she/he is allowed to refer to the corresponding reference translation (*ref*), with an advice that it is not only the correct translation.

2.5 Manual post-editing of MT output (*pe*)

Human workers were asked to post-edit MT outputs (*hyp*), i.e., to produce *pe*, under the following guidance.

(1) Refer only to *src* and *hyp* basically. Refer also to *ref* if necessary.

(2) Make each *hyp* grammatical and semantically appropriate with respect to its *src*, i.e., the quality of *pe* must be "A" or "S" in Table 2.

(3) Perform minimal edits, as we use *pe* for the reference of computing HTER (Snover et al., 2006).

The workers were also informed that we consider the following four edit operations equally.

Deletion of a word: Delete an unnecessary word: e.g., "the an" → "the"

Insertion of a word: Insert a missing but necessary word: e.g., "We will stay at hotel." → "We will stay at the hotel."

Substitution of a word: Substitute a word with another word. Change of inflection and conjugation is also regarded as this operation: e.g., "Can you teach me the way to the station?" → "Can you tell me the way to the station?"

Shift of a word or a phrase: Change the word order by moving a single word or a sequence of consecutive words: e.g., "I'll send a card my friend." → "I'll send my friend a card."[7]

2.6 Consistency check

Note that the last two tasks, i.e., grading and post-editing of MT outputs, were performed completely separately. Now, discrepancies between *grade* and *pe* were resolved in this final step. When both *grade* and *pe* for the same pair of *src* and *hyp* were registered, we assessed them according to the following three criteria.

[6] Their "AA" and "F" correspond to our "S" and "D," respectively.

[7] One can edit this *hyp* to "I'll send a card to my friend." In this case, the operation is considered as an "Insertion of a word (to)."

Table 3: Distribution of segments according to their grade.

Grade	*travel* (8,783 segments)						*hospital* (1,676 segments)					
	Ja→En		Ja→Zh		Ja→Ko		Ja→En		Ja→Zh		Ja→Ko	
	#seg	%	#seg	%	#seg	%	#seg	%	#seg	%	#seg	%
S	1,961	22.3%	2,827	32.2%	3,466	39.5%	95	5.7%	708	42.2%	903	53.9%
A	1,462	16.6%	1,874	21.3%	2,326	26.5%	107	6.4%	514	30.7%	482	28.8%
B	1,269	14.4%	1,275	14.5%	1,360	15.5%	181	10.8%	172	10.3%	166	9.9%
C	1,067	12.1%	899	10.2%	724	8.2%	333	19.9%	107	6.4%	97	5.8%
D	3,024	34.4%	1,908	21.7%	907	10.3%	960	57.3%	175	10.4%	28	1.7%

Table 4: Proximity between translations obtained through different ways.

Domain	Translations compared	BLEU (↑)			TER (↓)		
		Ja→En	Ja→Zh	Ja→Ko	Ja→En	Ja→Zh	Ja→Ko
travel	(a) *hyp* against *ref*	21.52	26.18	38.85	57.95	50.81	43.43
	(b) *hyp* against *pe*	51.97	69.44	81.98	35.14	19.20	12.25
	(c) *pe* against *ref*	49.00	39.73	49.11	34.46	38.79	34.75
hospital	(a) *hyp* against *ref*	9.19	30.38	51.01	75.35	48.54	32.44
	(b) *hyp* against *pe*	18.95	86.45	93.52	66.03	8.63	4.12
	(c) *pe* against *ref*	65.15	34.29	54.16	24.69	43.78	30.00

- If the *grade* is either "S" or "A" but *pe* is not identical to the given *hyp*, both grading and post-editing are performed again.
- If the *grade* is either "B," "C," or "D" but *pe* is identical to *hyp*, both grading and post-editing are performed again.
- If *hyp* is closer to *ref* than to *pe*, i.e., $\text{TER}(hyp, pe) > \text{TER}(hyp, ref)$, the number of edits is not minimal;[8] so post-editing is performed again.[9]

As there could be a variety of translation options, seeking the complete minimality does not seem feasible. Nevertheless, we introduced the last constraint, because we need less-edited translations as *pe*. To compute TER scores using TERCOM,[10] we tokenized *hyp*, *ref*, and *pe*, using the tool in Moses[11] for English MeCab[12] with mecab-ko-dic[13] for Korean. For Chinese, we regarded each character as one token.

3 Analyses of the Created Datasets

This section describes characteristics of the created datasets, observations, and remaining issues.

First, the results of manual grading are summarized in Table 3. While MT outputs for the *travel* domain were much better than the *hospital* domain in the Ja→En task, the segments in the *hospital* domain were better translated by the Ja→Zh and Ja→Ko MT systems.

Table 4 shows proximity in terms of BLEU (Papineni et al., 2002) and TER (Snover et al., 2006), between translations obtained through different ways. (a) "*hyp* against *ref*" presents what is measured in standard evaluation of MT outputs. The scores in these rows reflect the distribution of MT outputs shown in Table 3. On the other hand, (b) "*hyp* against *pe*" gauges the amount of post-edits. As we asked to perform only necessary edits to assure at least grade "A," the scores in these rows should be good in general. Only the exception is the *hospital* domain in the Ja→En task. As most of the MT outputs were of low quality, the workers tended to abandon them rather than correcting them. Finally, (c) "*pe* against *ref*" rows demonstrate that these two types of translations were not necessarily highly similar. Nevertheless, *pe* were certainly better than *hyp* with respect to *ref*. Again, *pe* in the *hospital* domain in the Ja→En task show exceptionally good scores. We plan to make an in-depth analysis with this respect.

The human judgment and the quantity of post-edits (HTER) evaluate the translation quality from different aspects. Indeed, as illustrated in Figure 2, many *hyp* that got grade "B" did not have smaller HTER score than those of grade "D." Figure 3 exemplifies some discrepancies between *grade* and HTER score observed in the Ja→En dataset. The

[8] This constraint can easily be satisfied by just copying *ref* to *pe*, but we prohibited this.

[9] We asked to restart from *hyp*, because resuming from the submitted *pe* would make the total number of edits unclear.

[10] http://www.cs.umd.edu/~snover/tercom/, version 0.7.25

[11] http://statmt.org/moses/, RELEASE-2.1.1

[12] https://github.com/taku910/mecab/, version 0.996

[13] https://bitbucket.org/eunjeon/mecab-ko-dic/, version 2.0.1-20150920

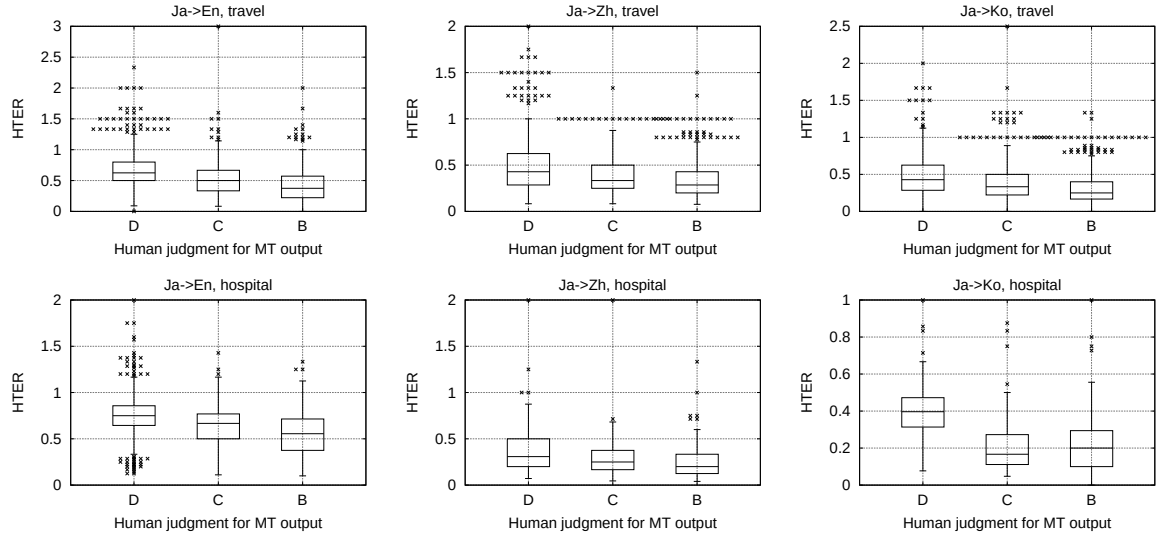

Figure 2: Distribution of sentence-wise HTER score with regard to each human judgment.

#1	*src*	多額の現金は持ってこないでください。
	ref	Please don't bring a lot of cash.
	hyp	Please bring a lot of cash.
	grade	D
	pe	Please don't bring a lot of cash.
	HTER	0.22
#2	*src*	首が痛くありませんか。
	ref	Doesn't your neck hurt?
	hyp	Do you have pain in my neck?
	grade	D
	pe	Do you have pain in your neck?
	HTER	0.13
#3	*src*	素晴らしい景色だね
	ref	It's a wonderful view, isn't it?
	hyp	It's beautiful scenery.
	grade	B
	pe	The scenery's beautiful, isn't it ?
	HTER	0.78

Figure 3: Examples from the Ja→En dataset.

hyp in the first two examples were graded "D," while they were only slightly edited. The *hyp* in #1 failed to appropriately convey the meaning of negation. On the other hand, considering that the segment #2 is given by a health care provider, the possessor of "首 (neck)" must not be him/her (the utterer), but the patient (the hearer). In both cases, the error in *hyp* is critical, even though it can be corrected with a small number of edits. This suggests that sentence-level QE systems should be optimized according to appropriate criteria, depending on their application.

There were also several examples that were graded "B" but were post-edited significantly. For instance, the *hyp* in #3 could be corrected by simply replacing the full stop with a tag question, i.e., "isn't it?" with a HTER score of 0.56. However, the worker also changed the syntactic structure of the main clause, increasing the HTER score. To avoid this kind of over-editing, the instruction in Section 2.5 should be improved.

Table 5: Number of segments in each partition.

Partition	*travel*	*hospital*	Merger
train	7,083	1,376	8,459
dev	850	150	1,000
test	850	150	1,000

4 Benchmarking

Using the created datasets, we conducted benchmarking experiments on word-level QE, sentence-level QE, and APE.

4.1 Common Settings

First, each of the *travel* and *hospital* datasets was randomly partitioned into training, development, and test sets as shown in Table 5. Although we believe that our datasets are useful for examining domain adaptation methods, in this paper, we report on experiments using the merger of data in the two domains. Table 6 summarizes the statistics of each partition in each task.[14] "BAD%-WQE" indicates the percentages of "BAD" tags for word-level QE (see Section 4.2 for details), while "BAD%-SQE" indicates the ratio of *hyp* that need post-editing, i.e., those graded either "B," "C," or "D."

[14] We tokenized them with our in-house tokenizer, which is also used in our speech translation service.

Table 6: Statistics of the training, development, and test partitions of the datasets.

Task	Partition	#seg	Tokens			Types			BAD%	
			src	*hyp*	*pe*	*src*	*hyp*	*pe*	WQE	SQE
Ja→En	train	8,459	65,855	59,377	63,970	5,739	3,772	4,475	29.0	65.2
	dev	1,000	7,657	7,004	7,526	1,680	1,201	1,365	28.9	66.9
	test	1,000	7,700	7,002	7,544	1,726	1,231	1,439	29.2	65.1
Ja→Zh	train	8,459	65,855	50,482	51,735	5,739	4,907	5,139	9.0	43.3
	dev	1,000	7,657	5,883	5,993	1,680	1,483	1,530	9.6	42.3
	test	1,000	7,700	5,915	6,042	1,726	1,516	1,562	9.9	44.9
Ja→Ko	train	8,459	65,844	65,520	66,550	5,739	5,103	5,213	7.6	31.3
	dev	1,000	7,657	7,674	7,791	1,680	1,598	1,632	8.3	32.2
	test	1,000	7,700	7,614	7,740	1,726	1,632	1,680	7.3	30.9

Table 7: Statistics of the DLC corpus.

Partition	#seg	Tokens				Types			
		Ja	En	Zh	Ko	Ja	En	Zh	Ko
train	1.57M	25.1M	22.3M	20.1M	24.0M	274,746	227,033	236,410	264,328
dev	14k	224k	200k	179k	215k	14,388	12,492	12,552	11,966

For the QE/APE tasks, due to the scarcity of training data, even baseline approaches have employed external resources, such as parallel and monolingual corpora, in addition to the task-specific training data. However, there is no publicly available parallel and monolingual data of spoken language in the language pairs of our concern. Therefore, we reluctantly employed an in-house parallel corpus of daily life conversations (DLC). Its statistics are shown in Table 7.

4.2 Word-level QE (WQE)

Given a pair of source text (*src*) and MT output (*hyp*), the task of word-level QE is to predict a sequence of tags with the same length as *hyp*, where each tag indicates how good the corresponding word in *hyp* is. While some previous studies, such as Bach et al. (2011), addressed to gauge the quality of each word with a real-valued score, WMT adopted a coarse-grained binary tag, i.e., {OK, BAD}, presumably because this form of tags can be automatically generated as the by-product of computing HTER score by comparing *hyp* with its post-edited version (*pe*) (Bojar et al., 2015). Following the recent convention in WMT, we automatically generated a sequence of binary tags for each pair of *src* and *hyp* using TERCOM. As the evaluation metrics, we used F_1 score of detecting "OK" tags (F_1-OK), that for "BAD" tags (F_1-BAD), and their product (F_1-mult) as in Bojar et al. (2016).

As a system for WQE, we adopted an implementation[15] based on a feed-forward neural network with its default setting. Following the investigation in Liu et al. (2017), we also generated a set of pseudo training data using the DLC corpus as follows.

Table 8: Pseudo data for the WQE task.

Task	Tokens	BAD%
Ja→En	10,945,486	50.3
Ja→Zh	9,867,440	39.4
Ja→Ko	11,891,369	30.6

Step 1. Phrase-based statistical MT systems for Ja→∗ translation tasks were built from the first half of the DLC corpus using Moses.

Step 2. Japanese sentences in the remaining half of the DLC corpus were decoded by the MT systems.

Step 3. Tag sequences for the MT outputs were given in the same manner as the manually created data, except that we regarded reference translations in the second half of the DLC corpus as post-edited MT outputs.

As presented in Table 8, we generated much larger data than the manually created training data in Table 6, although the pseudo training data tended to contain more "BAD" tags than the manually created data due to the independence between *hyp* and *ref*.

Our experimental results are presented in Table 9. The results for the Ja→En and Ja→Zh tasks are consistent to the observations in Liu et al. (2017), i.e., pseudo training data improve F_1-BAD scores. However, introduction of such data do not improve F_1-BAD in the Ja→Ko task.

[15] https://github.com/lemaoliu/qenn/

Table 9: Results for the WQE task.

System	F_1-mult (↑)			F_1-BAD (↑)			F_1-OK (↑)		
	Ja→En	Ja→Zh	Ja→Ko	Ja→En	Ja→Zh	Ja→Ko	Ja→En	Ja→Zh	Ja→Ko
All BAD	-	-	-	0.452	0.181	0.136	-	-	-
All OK	-	-	-	-	-	-	0.829	0.948	0.962
FNN-manual	0.345	0.205	0.295	0.469	0.229	0.313	0.736	0.896	0.942
FNN-pseudo	0.315	0.195	0.181	0.477	0.247	0.220	0.660	0.790	0.827
FNN-both	0.341	0.211	0.196	0.487	0.256	0.232	0.701	0.825	0.846

Table 10: Results for the SQE prediction task ("#f" indicates the number of features).

System	#f	Pearson's r (↑)			MAE (↓)			RMSE (↓)		
		Ja→En	Ja→Zh	Ja→Ko	Ja→En	Ja→Zh	Ja→Ko	Ja→En	Ja→Zh	Ja→Ko
Avg. of train	-	-	-	-	0.306	0.198	0.158	0.347	0.238	0.205
QuEst17	17	0.427	0.125	0.239	0.255	0.185	0.159	0.325	0.242	0.201
QuEst17+SntEmb	617	0.516	0.301	0.413	0.239	0.184	0.153	0.298	0.228	0.192

4.3 Sentence-level QE (SQE)

Given a pair of source text (*src*) and MT output (*hyp*), the task of sentence-level QE is to predict how good the entire *hyp* is, with respect to *src*. We conducted experiments on both of the HTER prediction and binary classification tasks.

4.3.1 Prediction of HTER

In WMT, this task is to predict the HTER score, directly from (*src*, *hyp*) pair (Specia et al., 2015), or indirectly through predicting the necessary edits in a similar manner to WQE (Kim and Lee, 2016).

We implemented a tool to extract a set of 17 features[16] of QuEst++ (Specia et al., 2015), which is regarded as the baseline of this task. To compute the features based on language models, we used the corresponding part of the DLC corpus. To estimate the translation-related features, such as the number of translations per word in *src*, we trained a phrase-table on the DLC corpus using Moses. Following the findings in Shah et al. (2016), we also incorporated the distributed representations of *src* and *hyp*. First, word embeddings with 300 dimensions were learned from each part of the DLC corpus using word2vec[17] with its default parameters. Then, the embedding for a given segment is computed by averaging the embeddings of its constituent words, assuming the additive compositionality (Mikolov et al., 2013). During the computation, unknown words were mapped to a zero vector. Finally, values for each of 300 dimensions were regarded as additional features.

The extracted features were used to train support vector regression (SVR) models with a radial basis function (RBF) kernel.[18] Hyper-parameters were optimized with respect to the development set, through a grid search to maximize the Pearson's correlation coefficient r between the predicted HTER and the gold HTER.

Table 10 justifies that sentence embeddings obtained by such a naive way[19] can improve the performance of predicting HTER score, irrespective of the evaluation metrics: Pearson's correlation coefficient r, mean average error (MAE), and root mean squared error (RMSE).

4.3.2 Binary Classification

We assume that users of speech translation services are usually not competent in the target language. Thus, when we consider directly delivering the MT outputs to such users, their quality in terms of our *grade* seems more intuitive than HTER.

We evaluated how well the same feature sets in Section 4.3.1 can predict the grade, using support vector classifier (SVC) instead of SVR. Hyper-parameters were optimized so that they maximize F_1-mult on the development set. The systems (feature sets) were evaluated with the same metrics as in WQE.

As presented in Table 11, we obtained consistent results that baseline systems with QuEst++ features can be improved by incorporating the distributed representations of *src* and *hyp*.

[16] http://www.quest.dcs.shef.ac.uk/quest_files/features_blackbox_baseline_17

[17] https://github.com/tmikolov/word2vec/

[18] http://chasen.org/~taku/software/TinySVM/

[19] As a more advanced alternative, one can train a neural MT system and retrieve annotations from RNN's hidden states as proposed in (Kim and Lee, 2016).

Table 11: Results for the SQE classification task ("#f" indicates the number of features).

System	#f	F_1-mult (↑)			F_1-BAD (↑)			F_1-OK (↑)		
		Ja→En	Ja→Zh	Ja→Ko	Ja→En	Ja→Zh	Ja→Ko	Ja→En	Ja→Zh	Ja→Ko
All BAD	-	-	-	-	0.789	0.620	0.472	-	-	-
All OK	-	-	-	-	-	-	-	0.517	0.711	0.817
QuEst17	17	0.335	0.295	0.310	0.765	0.442	0.403	0.438	0.667	0.770
QuEst17+SntEmb	617	0.450	0.410	0.396	0.798	0.584	0.480	0.563	0.702	0.825

Table 12: Results for the APE task.

Method	BLEU (↑)			TER (↓)		
	Ja→En	Ja→Zh	Ja→Ko	Ja→En	Ja→Zh	Ja→Ko
Raw MT output	43.74	73.14	85.52	42.21	16.98	9.87
(a) APE w/ gold data only	43.38	72.28	84.87	42.33	17.53	10.31
(b) (a) + bitext back-off	44.00	73.01	85.53	41.87	17.05	9.87
(c) (b) + pseudo training data	43.90	73.15	85.57	41.95	16.97	9.82

4.4 APE

The task of APE is to automatically post-edit MT outputs (*hyp*). Although there are a number of methods that also refer to *src* (Béchara et al., 2011; Junczys-Dowmunt and Grundkiewicz, 2016), we have so far examined only classic baseline methods based on phrase-based statistical MT.

The first system (a) was trained only on the gold data (Simard et al., 2007a) using Moses. However, this system tended to deteriorate the translation quality in terms of BLUE and TER, presumably due to the scarcity of training data. Then, our second model (b) introduced identical pairs of sentences in the target side of our DLC corpus in order to conservatively retain grammatical fragment within *hyp*. By (re-)decoding the *hyp* using the multiple decoding path ability of Moses,[20] this model significantly improved the naive baseline system (a), but the translation quality was not consistently better depending on the language pair.

Finally, we introduced in the third system (c) yet another phrase table learned from pseudo training data as proposed by Simard et al. (2007b). Our pseudo training data were obtained in the same manner as those for WQE (see Section 4.2); we coupled each of the decoded result to its corresponding reference translation in the DLC corpus. As summarized in Table 12, this model led to a slight but consistent improvement on both metrics in the all tasks.

[20]We used the "either" strategy. If a phrase pair appears in more than one phrase table, different decoding paths are generated and each considers only the corresponding features for scoring.

5 Conclusion

Aiming to promote the research on quality estimation (QE) and automatic post-editing (APE) of MT outputs, especially for those among Asian languages, we have created new datasets for the Japanese to English, Chinese, and Korean translation tasks. This paper described the process of corpus creation and observations from the created datasets. We also presented our benchmarking experiments using the created datasets, for all of the tasks in our concern: word-level QE, two variants of sentence-level QE, and APE. Although the methods examined in this paper could be far from the state-of-the-art, we confirmed that the performance of these tasks can be improved by introducing features and pseudo training data that had been proven useful in the literature.

Following the emergence of neural MT, we are now working on extending the datasets with translations of such systems. We are planning to further improve the performance on the QE/APE tasks, and to investigate applications of the technologies, including enhancing the functionality of our speech translation service, and filtering automatically harvested parallel sentences (Sennrich et al., 2016; Marie and Fujita, 2017).

Acknowledgments

We are deeply grateful to the anonymous reviewers for their thorough and constructive comments. This work was conducted under the program "Promotion of Global Communications Plan: Research, Development, and Social Demonstration of Multilingual Speech Translation Technology" of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: A method for measuring machine translation confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 211–219.

Yehosua Bar-Hillel. 1951. The present state of research on mechanical translation. *American Documentation*, 2(4):229–237.

Hanna Béchara, Yanjun Ma, and Josef van Genabith. 2011. Statistical post-editing for a statistical MT system. In *Proceedings of the 13th Machine Translation Summit* (*MT Summit XIII*), pages 308–315.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the 9th Workshop on Statistical Machine Translation* (*WMT*), pages 12–58.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the 10th Workshop on Statistical Machine Translation* (*WMT*), pages 1–46.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the 1st Conference on Machine Translation* (*WMT*), pages 131–198.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation. In *Proceedings of the 2nd Conference on Machine Translation* (*WMT*).

Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. 2013. Overview of the patent machine translation task at the NTCIR-10 workshop. In *Proceedings of NTCIR-10 Workshop Meeting*, pages 260–286.

W. John Hutchins and Harold L. Somers. 1992. *An Introduction to Machine Translation*. Academic Press.

ISO/TC27. 2017. ISO 18587:2017 translation services: Post-editing of machine translation output: Requirements.

Japan National Tourism Organization. 2017. Foreign visitors & Japanese departures. https://www.jnto.go.jp/eng/ttp/sta/.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the 1st Conference on Machine Translation* (*WMT*), pages 751–758.

Hyun Kim and Jong-Hyeok Lee. 2016. A recurrent neural networks approach for estimating the quality of machine translation output. In *Proceedings of Human Language Technologies: The 2016 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), pages 494–498.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press.

Lemao Liu, Atsushi Fujita, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2017. Translation quality estimation using only bilingual corpora. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (*TASLP*), 25(9):1458–1468.

Benjamin Marie and Atsushi Fujita. 2017. Efficient extraction of pseudo-parallel sentences from raw monolingual data using word embeddings. In *Proceedings of the 55nd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 392–398.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems* (*NIPS*), pages 3111–3119.

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* (*ACL*), pages 311–318.

Marion Potet, Emmanuelle Esperança-Rodier, Laurent Besacier, and Hervé Blanchon. 2012. Collection of a large database of French–English SMT output corrections. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (*LREC*).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54nd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 86–96.

Kashif Shah, Fethi Bougares, Loïc Barrault, and Lucia Specia. 2016. SHEF-LIUM-NN: Sentence-level quality estimation with neural network features. In *Proceedings of the 1st Conference on Machine Translation* (*WMT*), pages 838–842.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007a. Statistical phrase-based post-editing. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), pages 508–515.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007b. Rule-based translation with statistical phrase-basd post-editing. In *Proceedings of the 2nd Workshop on Statistical Machine Translation* (*WMT*), pages 203–206.

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 Biennial Conference of the Association for Machine Translation in the Americas* (*AMTA*), pages 223–231.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with QuEst++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.