Findings of the WMT 2024 Shared Task on Non-Repetitive Translation

Kazutaka Kinugawa[1], Hideya Mino[1], Isao Goto[2], Naoto Shirai[1]

[1]NHK Science & Technology Research Laboratories, Tokyo, Japan
[2]Ehime University, Ehime, Japan
{kinugawa.k-jg,mino.h-gq,shirai.n-hk}@nhk.or.jp
goto.isao.fn@ehime-u.ac.jp

Abstract

The repetition of words in an English sentence can create a monotonous or awkward impression. In such cases, repetition should be avoided appropriately. To evaluate the performance of machine translation (MT) systems in avoiding such repetition and outputting more polished translations, we presented the shared task of controlling the lexical choice of MT systems. From Japanese–English parallel news articles, we collected several hundred sentence pairs in which the source sentences containing repeated words were translated in a style that avoided repetition. Participants were required to encourage the MT system to output tokens in a *non-repetitive* manner while maintaining translation quality. We conducted human and automatic evaluations of systems submitted by two teams based on an encoder-decoder Transformer and a large language model, respectively. From the experimental results and analysis, we report a series of findings on this task.

1 Introduction

The development of neural models has improved the performance of machine translation (MT) significantly (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). MT systems are now used in a variety of real-world scenarios; however, challenges remain for such systems that assist human writers. Specifically, the MT output must not only be *adequate* and *fluent* but also follow the writing style of the target domain. For example, it is advisable for an application in the English news domain to follow rules such as the use of active rather than passive voice, the use of the affirmative rather than the negative, and the avoidance of redundant phrases (Block, 1994; Cappon, 2019; Papper, 2021). Among these writing style rules, we focus on the rule regarding the repetition of words in the English news domain. Generally, common words repeated in a sentence can create a monotonous or awkward impression, and in such cases, repetition should be avoided appropriately (Burstein and Wolska, 2003). Typical workarounds are (1) the removal of redundant terms, if possible (Strunk and White, 1999) or (2) the use of alternative words, such as synonyms, as substitutes.[1] In this paper, we refer to translation techniques (1) and (2) as *reduction* and *substitution*, respectively, and call the translation style using these techniques a *non-repetitive style*. Figure 1 shows an example of a non-repetitive style translation from a Japanese–English parallel news article. We observe that human writers in the English news domain often translate Japanese text with such reduction and substitution.[2] These translation techniques arise from the difference between the styles of the source and

Ja: ...入学予定者７人が教育方針や私立小への入学などを理由に入学を辞退した。
En(trans): ..., seven students scheduled to enroll withdrew their enrollment due to reasons such as the educational policy and enrolling in a private school.
En: ..., seven children dropped plans to enter the school, with parents citing disagreements with its education policy, decisions to join private schools or other reasons, ...

Figure 1: Motivating example from a Japanese–English parallel news article along with a consistent translation ("En(trans)") for comparison. Repeated words and their counterparts are highlighted. "入学" is intentionally removed (*reduction*), probably because it is contextually obvious. In this paper, we distinguish this type of removal from undertranslation. Additionally, "入学" and "入学" are translated differently as "join" and "enter," respectively (*substitution*).

[1]https://effectiviology.com/writing-tips-from-the-elements-of-style/#Avoid_repetition

[2]Other examples are listed in Appendix A.

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 715–727
November 15–16, 2024.

target languages; that is, an article in the source language was originally produced by a writer who attached importance to conveying content without the reader misunderstanding it by using the same expressions consistently, and then it was translated into the target language by a writer who was encouraged to (or preferred to) translate it in a more diverse or concise way. The assumption in this task is that sentences translated in a simple word-by-word manner cannot be suited to the target domain. We could thus associate these translation techniques with a type of rewriting. Although this task focuses on the news domain, the monotony or awkwardness arising from the repetition of words in English is also a common problem in other domains.

Given this motivation, we presented a shared task for non-repetitive translation. To configure appropriate settings, we limited the task to one-to-one or two-to-two translations. We hypothesized that the closer the distance between repeated words, the greater the need to translate using reduction or substitution. Additionally, we targeted the repetition of common words because such words tend to be substituted according to the findings of Guillou (2013). We qualitatively categorized several patterns of non-repetitive style translations, and then collected several hundred instances in which some words were repeated in the source sentence and translated using reduction or substitution, which we used as development and test data. In the remainder of this paper, we first explain the research background of this task (§2). Next, we describe the task definition (§3), dataset we prepared (§4), evaluation methods (§5), and submitted systems (§6). Finally, we present the results and some analysis (§7).

2 Related Work

Contrast with Consistent Translation In the context of MT research, lexically consistent translation (in this paper, we also refer to this as *repetitive style translation*) has been studied actively (Pu et al., 2017; Kuang et al., 2018; Tu et al., 2018; Lyu et al., 2021, 2022). A representative study is the hypothesis of "one translation per discourse," which was advocated by Carpuat (2009). The motivation for these studies is the assumption that translating text in a consistent style should be encouraged because this style is unambiguous and accurate for readers. Moreover, from the viewpoint of experimental evaluation, many researchers have reported that BLEU scores improved as a result of encouraging consistent translation (Lyu et al., 2021, 2022). However, it is debatable whether all words should be translated consistently. Translation consistency can depend on several factors, such as the target domain, type of words, and translation direction (Guillou, 2013). For example, it is indisputable that technical terms in the patent domain should be translated consistently. By contrast, Guillou (2013) reported that high-frequency verbs are often translated in diverse ways in English–French translation. While improving document-level consistency based on the postprocess approach, Zhang et al. (2023) also mentioned the side effect of the loss of translation diversity. From another point of view, consistent translation has the risk of leading to a robotic wording and giving a monotonous or awkward impression to readers, as shown in Figure 1. By contrast, Cappon (2019) claimed that excessive substitution may obscure the meaning of the sentence. In monolingual writing, this phenomenon is derided as *the elegant variation*.[3] To summarize, there is a trade-off between ambiguity and monotony. This task particularly focuses on the latter aspect, which has not often been addressed in previous studies. To the best of our knowledge, no test sets exist for directly evaluating such a translation style.

Reduction and Substitution Although several studies have been conducted related to non-repetitive translation, the scope of our research is different. First, several researchers have addressed the problem of controlling the output length of MT systems (Lakew et al., 2019; Schioppa et al., 2021). Typically, special tokens representing the output length at several discrete levels are inserted into source sentences. Although this approach is associated with reduction, our task requires a more meticulous omission of specific words in sentences. Regarding substitutions, MT systems are sometimes required to select infrequent words from the vocabulary. However, researchers have reported that MT systems are biased toward outputting high-frequency target words (Ott et al., 2018; Gu et al., 2020) and tend to produce lexically poorer translations than humans (Vanmassenhove et al., 2019, 2021). Gu et al. (2020) designed the objective function so that low-frequency target tokens were more likely to be output. However, they conducted

[3] https://en.wikipedia.org/wiki/Elegant_variation

the experiment using regular corpora and did not present a perspective on in what scenarios low-frequency words should be output. By contrast, we set up a more specific scenario.

This task is also related to research outside of translation technique. Neural models have the traditional problem of not outputting the end-of-sequence token while generating the same tokens endlessly. To alleviate this problem, several approaches including learning-based methods (Welleck et al., 2020) and decoding-based methods (Keskar et al., 2019), have been proposed. Although the goal is different, these studies are also relevant to our task in the sense that word repetition should be avoided.

3 Task Definition

Our task focused on lexical choice in MT, particularly choice regarding repeated words in a source sentence. The translation direction was Japanese to English. Participants were required to control an MT system using reduction or substitution so that it did not output the same words for certain repeated words in a source sentence. Simultaneously, participants also needed to maintain translation quality as much as possible.

The challenges underlying this task included the following:

- Maintaining the balance between translation quality and controlling the output: Translation quality can be degraded when the non-repetitive style is enforced inappropriately.

- Avoiding bias toward high-frequency bilingual word pairs: Generally, for a given source word, high-frequency target words associated with it are more likely to be output. This can make it difficult to determine appropriate substitutions for some words.

- Predicting which words can be reduced or substituted: It is not easy to make an appropriate prediction because it depends on the context within the sentence.

- Mining training instances: Translations with reduction can be particularly difficult to identify in noisy corpora because of the challenge of discriminating them from undertranslations.

4 Dataset

We prepared the training, development, and test data for this task. They were all sourced from Japanese–English news articles published by Jiji Press LTD., a Japanese news agency. We annotated the development and test data for this task, whereas the training data comprised a regular MT corpus.

4.1 Development and Test Data

We provided development and test sets for this task, which we refer to as Jiji 2023 data and Jiji 2024 data, respectively. These data included 162 and 479 instances, respectively. The Jiji 2023 data were originally built for the Non-Repetitive Translation Task in WAT 2023 (Nakazawa et al., 2023). We reviewed the data and filtered out some instances this year. By contrast, the Jiji 2024 data were newly created in this year. In both datasets, all Japanese sentences contained some repeated words that were translated into English with reduction or substitution. From Japanese–English news articles, we first automatically created sentence pairs based on lexical similarities using the method of (Utiyama and Isahara, 2007) and then manually selected appropriate instances. To reduce the negative effects of imbalanced content in the source and target sentences, the Japanese sentences in the Jiji 2023 and 2024 data were manually translated from English by professional translators while preserving as much of the vocabulary of the original Japanese sentences as possible. Both the released development and test sets contained raw and tagged parallel data. In the tagged data, we marked repeated words in the source sentence and their counterparts in the target sentence with tags, which indicated that these words were evaluation targets. Examples are shown in Table 1. The respective attributes inside the tags indicate the following:

id: This indicates the IDs of repeated words. In the example, two tagged repeated words are included, that is, “機能” (“id=0”) and “製品” (“id=1”). The number of instances including multiple tagged repeated words, such as this example, are limited. Additionally, the number of types of repeated words in one instance is one or two.

ref: This indicates the IDs of pairs of source words and their counterparts, such as (“製品,” “models”) (i.e., “id=1” and “ref=0”) and (“製品,” “products”) (i.e., “id=1” and “ref=1”).

Ja	JEMAの担当者は白物家電について、「<target id=0 ref=0 type=s>機能<\target>を絞った低価格<target id=1 ref=0 type=s>製品<\target>、高価格な高<target id=0 ref=1 type=s>機能<\target><target id=1 ref=1 type=s>製品<\target>とも好調だ」と述べている。
En	“Shipments have been robust for both low-priced <target id=1 ref=0 type=s>models<\target> with reduced <target id=0 ref=0 type=s>functions<\target> and expensive <target id=0 ref=1 type=s>high-spec<\target><target id=1 ref=1 type=s>products<\target>,” a JEMA official said.

Table 1: Examples of tagged instances in the development and test data. The tags are highlighted.

Split	# Parallel sentences
train	200K
dev	479
test	1851

Table 2: Statistics of the Jiji 2020 data. Note that “dev” and “test” in the table are different from the Jiji 2023 and 2024 data.

type: This indicates whether tagged source words are substituted (“s”) or reduced (“r”).

Note that not all words repeated in the source sentence were evaluation targets. This is because some words, such as proper nouns and technical terms, should be translated consistently, even if they were repeated in the sentence. We provided the tagged development data to help to tune the model during training. However, participants could not use the tagged test data when submitting the system results. In this task, the systems had to detect repeated words that could be reduced or substituted on their own.

4.2 Training Data

Regarding the training data, we provided all the data from the WAT 2020 Newswire tasks (Nakazawa et al., 2020), which were also constructed from Jiji news articles and have been continuously used in WAT since 2020 (Nakazawa et al., 2020, 2021, 2022, 2023). For simplicity, we refer to these data as Jiji 2020 data. The main files in the Jiji 2020 data are shown in Table 2. These data are a regular parallel corpus. They were not annotated specifically for this task but were in exactly the same domain as the Jiji 2023 and 2024 data. Although the development and test sets in the Jiji 2020 data, which are described as “dev” and “test” in Table 2, were not directly related to the evaluation of this task, they could be used to measure basic translation performance during training. Unfortunately, the number of parallel sentences in the Jiji 2020 data was limited. Thus, we allowed participants to use any other corpora for training.

5 Evaluation

We conducted both human and automatic evaluation. We based the main results of this task on the human evaluation and prepared the automatic evaluation as secondary metrics. Again, the goal of this task was to control an MT system to output translations in a non-repetitive style while maintaining translation quality.

5.1 Human Evaluation

We evaluated system performance using the total number of outputs that met both acceptable translation adequacy and appropriate lexical choice. Both aspects were checked by three human translators, who were assigned by the authors.

Translation Style Regarding the evaluation for lexical choice, the human translators checked whether the translations for the tagged source words were correctly written in a non-repetitive style. Whether untagged repeated words were translated in a repetitive or non-repetitive way did not affect this evaluation. Moreover, the technique (i.e., reduction or substitution) did not have to be consistent with that of the reference translation. In our preliminary investigations, we qualitatively studied the lexical choices of several translators, and observed cases in which one translator chose substitution, and another chose reduction. Additionally, the systems did not have to choose the same words used in the reference, provided the meaning was appropriate. The determination of substitution or repetition was essentially based on the word stem. For example, conversions between voice (e.g., “attack” and “be attacked”), tense (e.g., “study” and “studied”), and parts of speech (e.g., “problematic” and “problem”) were not considered to be substitutions. Conversions to idioms (e.g., “visit” and “pay a visit”) were an exception and handled as substitutions. This evaluation is not trivial. For example, it is difficult to establish uniform guidelines for determining the correctness of synonyms in substitution and whether they are appropriate reductions

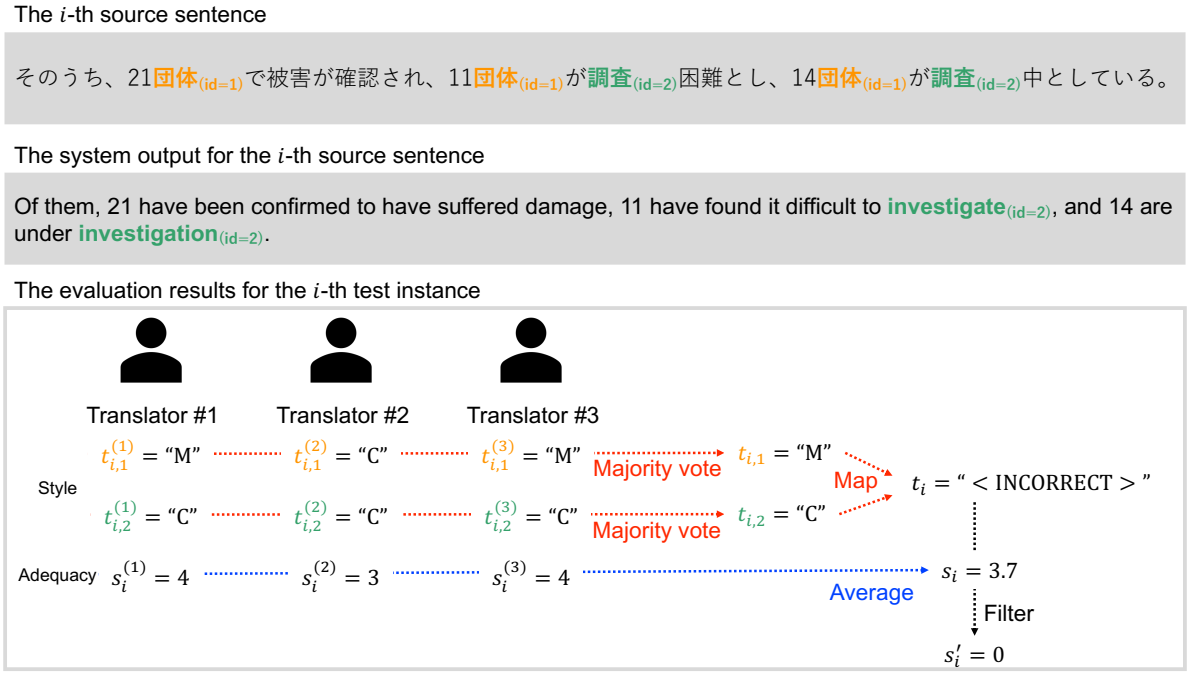

Figure 2: Example of human evaluation for the i-th test instance. "団体" (id=1) is undertranslated (at least one counterpart should appear in the output in this case) (the label is thus "M"), and "調査" (id=2) is translated in a repetitive style (the label is thus "C"). For simplicity, 1-indexed IDs are used for the repeated words.

or inappropriate omissions. Thus, we adopted a majority vote by the three human translators in this evaluation process.

Next, we explain the evaluation procedure for the i-th test instance, which is also illustrated in Figure 2. Other test instances were also evaluated in the same manner and all results were finally aggregated. First, each translator labeled the translations for the tagged source words in the i-th test instance as "S" (substitution), "R" (reduction), "C" (consistent, i.e., repetitive), or "M" (mistranslation or undertranslation). Note that "S," "R," and "C" implicitly indicate that the meaning of the translation is correct. Let the label for the j-th evaluation target in the i-th test instance given by the k-th translator be $t_{i,j}^{(k)}$. Next, the three labels $t_{i,j}^{(1)}$, $t_{i,j}^{(2)}$, and $t_{i,j}^{(3)}$ were reduced to one by a majority vote, which we denote by $t_{i,j}$. Because the number of types of labels was more than two, three labels could all be different. Although we assumed that such a case was limited, we introduced an additional heuristic rule to determine the label as follows:

- If the label set was equal to {"C,""R,""S"}, "S" was assigned to $t_{i,j}$: Because two translators thought it was correctly translated in a non-repetitive style, the label should be "R" or "S." Next, because two translators thought the word was not reduced, the label was determined to be "S."
- If the label set was equal to {"M,""R,""S"}, "R" was assigned to $t_{i,j}$: Because two translators thought it was correctly translated in a non-repetitive style, the label should be "R" or "S." Next, the label "M" was assigned probably because that translator thought some necessary word was omitted. Thus, the label was determined to be "S."
- If the label set was equal to {"M,""C,""S"}, "S" was assigned to $t_{i,j}$: Because two translators thought the meaning of the translation was correct, the label should be "C" or "S." Next, the label "M" was assigned probably because that translator thought some word had a slightly different nuance. Thus, the label was determined to be "S."
- If the label set was equal to {"M,""C,""R"}, "R" was assigned to $t_{i,j}$: Because two translators thought the meaning of the translation was correct, the label should be "C" or "S." Next, the label "M" was assigned probably because that translator thought some necessary word was omitted. Thus, the label was determined to be "R."

Finally, one representative label was assigned to the i-th test instance, which we denote by t_i. Representative labels were chosen from "<NON-REP>," "<REP>," and "<INCORRECT>." For test instances including only one target, the representative label t_i was simply mapped from $t_{i,1}$ as follows:

- If $t_{i,1}$ was "R" or "S," "<NON-REP>" was assigned to t_i.
- If $t_{i,1}$ was "M," "<INCORRECT>" was assigned to t_i.
- If $t_{i,1}$ was "C," "<REP>" was assigned to t_i.

For test instances including two targets, the representative label t_i was determined as follows:

- If $t_{i,1}$ was "R" or "S," and $t_{i,2}$ was "R" or "S," "<NON-REP>" was assigned to t_i.
- If $t_{i,1}$ was "M" or $t_{i,2}$ was "M," "<INCORRECT>" was assigned to t_i.
- Otherwise, "<REP>" was assigned to t_i.

Translation Accuracy In this task, the content of the system output may be omitted incorrectly or obscured if reduction or substitution is enforced inappropriately. Thus, we measured translation adequacy for system outputs. The evaluation framework was based on Japanese Patent Office (JPO) adequacy.[4] This criterion is well established and has also been used in domains other than patents.

Specifically, the k-th translator assigned a five-level discrete score $s_i^{(k)} \in \{1, 2, 3, 4, 5\}$ to the i-th system output. Next, we averaged $s_i^{(1)}$, $s_i^{(2)}$, and $s_i^{(3)}$ to s_i. Additionally, to view the balance between translation style and adequacy, we reflected the style label t_i in the adequacy score s_i. If the translation style was not "<NON-REP>," we reduced the adequacy score s_i to 0. We refer to this metric as *filtered adequacy* and denote it by s'_i.

5.2 Automatic Evaluation

We also automatically predicted whether the target word was translated in a repetitive style. Note that "<NON-REP>" and "<INCORRECT>" could not be discriminated in this process. Thus, we introduced one more label "<NOT-REP>," which indicated "<NON-REP>" or "<INCORRECT>."

[4] https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/tokkyohonyaku_hyouka.html (in Japanese)

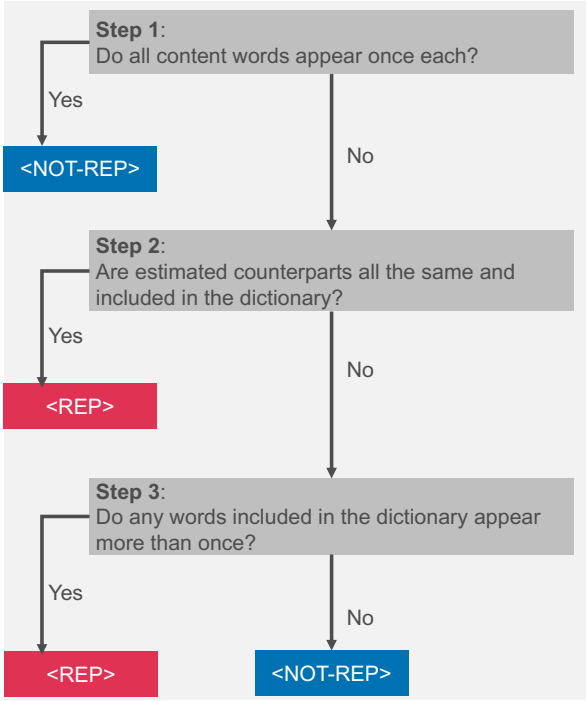

Figure 3: Yes/no flowchart for predicting translation styles.

As a preprocess, we built a bilingual dictionary from the Jiji 2020 data and JParaCrawl v3.0 (Morishita et al., 2022). We aggregated translations of evaluation target words in the Jiji 2024 data by running the AWESOME aligner (Dou and Neubig, 2021) on the above corpora. Let the j-th evaluation target word in the i-th source sentence be $w_{i,j}$. Based on the alignment results, we obtained a set of possible counterparts of $w_{i,j}$, which we denoted by $S_{w_{i,j}}$. We then removed low-frequency counterparts from $S_{w_{i,j}}$ to limit the maximum dictionary size $|S_{w_{i,j}}|$ to 10. We predicted a style label by applying several simple binary classifications in order of reliability confidence as follows:

(1) *Do all tokens appear once each?*: If all content words appear once each in the i-th system translation, this output is classified as "<NOT-REP>."

(2) *Are estimated counterparts all the same and included in the dictionary?*: First, we estimate counterparts of $w_{i,j}$ using the word aligner. If these counterparts are all the same and exist in $S_{w_{i,j}}$, this output is classified as "<REP>."

(3) *Do any tokens in the dictionary appear more than once?*: If any word in S_{w_x} appears more than once, this output is classified as a repetitive style; otherwise, the output is classified

as "<NOT-REP>."

We designed the third block to mitigate misclassification caused by alignment errors in (2). The above procedures are illustrated in Figure 3. Finally, we calculated the percentage of instances labeled as "<REP>" in the test set. We refer to this metric as *repetition rate*.

To measure translation quality, we also computed BLUE scores (Papineni et al., 2002) using SacreBleu (Post, 2018).[5]

6 Systems

In this shared task, two teams submitted the system and description paper. In this section, we provide an overview of the submitted systems and the baseline system that we built. For comparison, resources used by each system are listed in Table 3.

6.1 Baseline

As a baseline, we built an MT system using fairseq (Ott et al., 2019). We adopted Transformer (big) (Vaswani et al., 2017) as the architecture, and used the Jiji 2020 and JParaCrawl v3.0 (Morishita et al., 2022) as training data. We based the method on the tagging approach (Sennrich et al., 2016; Lakew et al., 2019; Johnson et al., 2017; Schioppa et al., 2021). Specifically, we introduced style and domain tags, and combined them. First, from the Jiji 2020 data and JParaCrawl v3.0, we mined sentence pairs in which some content words were repeated in the source sentence and no content words were repeated in the target sentence. We detected content words in Japanese and English sentences using GiNZA[6] and spaCy,[7] respectively. To avoid selecting noisy instances, we excluded parallel sentences with lexical similarity scores less than 0.7 from the tagging. Specifically, we prepended the style tag "<NON-REP>" and all repeated words to the source sentences as follows:

Src: <NON-REP> <文書> 米国立公文書館が文書を保管していた。

Second, we also attached the domain tag "<JIJI>" to training instances from the Jiji 2020 data. Similarly, we did not tag sentence pairs with lexical similarity scores less than 0.7. We prepended the domain tag to the source sentences as follows:

System	Resource
Baseline	JParaCrawl v3.0, Jiji 2020
SYSTRAN	all Ja–En data from OPUS, Jiji 2020
Waseda Riko	Claude 3.5 Sonnet, examples from the task website[8]

Table 3: Comparison of resources used by each system. Although the Waseda Riko system did not explicitly use the data on which Claude 3.5 Sonnet was built, they are also listed as "Claude 3.5 Sonnet" in the table.

Src: <JIJI> <NON-REP> <文書> 米国立公文書館が文書を保管していた。

For inference, we prepended the style and domain tags to all the test source sentences. In this system, we adopted the same hyperparameter settings as Morishita et al. (2022).

6.2 SYSTRAN (Avila and Crego, 2024)

The team introduced a *repetition penalty* in the fine-tuning phase. The method was inspired by label smoothing (Szegedy et al., 2015). For training instances including word repetition in the target sentence, the ground-truth score corresponding to the repeated word was decreased from 1. The team automatically detected such instances using the spaCy tokenizer[9] and GIZA++ toolkit (Och and Ney, 2003). Specifically, the repetition penalty was combined with label smoothing, and is formulated as follows:

$$q'_t = (1-\epsilon)(1-\alpha_t)q_t + \frac{\epsilon}{V},$$

where q_t indicates a one-hot vector used as the ground-truth label at the t-th time step, at which a repeated word appears, ϵ is a hyperparameter for label smoothing and V is the vocabulary size. α_t is also a hyperparameter used to control the degree to which word repetition is discouraged. The team first trained a Transformer encoder-decoder model on parallel sentences from OPUS[10] and then fine-tuned the model on parallel sentences from the Jiji 2020 data using the above technique. To avoid feeding noisy instances into the model, the team used back-translated sentences instead of the original sentence pairs in the fine-tuning stage.

6.3 Waseda Riko (Wang et al., 2024)

The team built a large language model (LLM)-based pipeline. The procedure was composed of

[5] nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.2

[6] https://megagonlabs.github.io/ginza/

[7] https://spacy.io/usage

[8] See Appendix A. These data were used for few-shot prompts.

[9] https://spacy.io/usage

[10] https://opus.nlpl.eu/

System	Adequacy (↑)	Translation Style			Filtered Adequacy (↑)
		% <NON-REP>	% <REP>	% <INCORRECT>	
Waseda Riko	**4.6**	**89.8**	8.1	2.1	**4.1**
SYSTRAN	3.9	32.3	53.8	13.6	1.3
Baseline	3.9	50.2	27.4	22.3	2.1

Table 4: Human evaluation results.

System	BLEU (%) (↑)	Translation Style		Repetition Rate (%) (↓)
		# <NOT-REP>	# <REP>	
Waseda Riko	24.4	413	57	**12.1**
SYSTRAN	28.9	214	256	54.5
Baseline	**29.1**	332	138	29.4

Table 5: Automatic evaluation results.

the following four steps:

(1) Preprocess: Detect repeated words from the source sentence using the MeCab tokenizer (Kudo et al., 2004) and tag these possible repeated words.

(2) Translation: Instruct the LLM to translate the tagged source sentence in a non-repetitive manner using a few-shot prompt (Brown et al., 2020).

(3) Proofreading: Instruct the LLM to review the output in the previous step and rewrite the translation as needed to enhance the result.

(4) Postprocess: Tag the counterparts in the target sentence.[11]

The team used Claude 3.5 Sonnet[12] and designed a prompt suited for this task. Specifically, they instructed the LLM to output translations along with the estimated counterparts and translation labels in JSON format. Because of this structured output design, the following processes were performed successfully.

7 Results and Discussion

7.1 Human Evaluation

We summarize the human evaluation scores of all systems in Table 4.[13] The Waseda Riko system achieved the best results in both translation adequacy and style control. Focusing on the drop from the adequacy score to the filtered adequacy score, the baseline system lost 1.8 points, whereas the Waseda Riko system only decreased by 0.5 points. This difference highlights that the Waseda Riko team successfully controlled the translation style without compromising translation quality. The SYSTRAN system achieved an adequacy score competitive with that of the baseline system, but passed more source sentences in a repetitive style. By contrast, the baseline system was the worst in terms of the percentage of incorrect instances. Considering the difference between the SYSTRAN and baseline systems, a trade-off existed between style control and translation adequacy.

The basic idea of the Waseda Riko system is similar to that of the baseline system: possible repeated words in the source sentence were automatically detected using a third-party tokenizer and the model was explicitly informed about them. (Wang et al. (2024) also reported that it was still difficult for LLMs to consistently identify repeated words in the input sentence.) Although the baseline system was trained on parallel sentences that were (possibly) translated in a non-repetitive style, the percentage of test instances in the desired style was 50%. Although the results of the Waseda Riko team were also supported by the high performance of the commercial LLM, their proposed prompt design and pipeline configuration were equally important. The key was how to provide the instruction to "translate in a non-repetitive style," which is (probably) new and complex for many LLMs. We attempted to instruct GPT-3.5 turbo[14] to solve this task using a simple prompt, such as *"Translate the following Japanese news text into English using as few of the same content words as possible,"* in our preliminary experiments, but this did not work well.

[11] Human evaluation was performed on untagged translations; thus, it was not necessary to tag the system output.

[12] https://www.anthropic.com/claude

[13] Detailed statistics are listed in Appendix B.

[14] https://azure.microsoft.com/en-us/products/ai-services/openai-service

7.2 Automatic Evaluation

We also summarize the automatic evaluation scores of all systems in Table 5. In contrast to the human evaluation, the baseline and SYSTRAN systems achieved a better BLEU score than the Waseda Riko system. This gap depended on whether the systems used the Jiji 2020 data for training. Although the Waseda Riko team analyzed these data and then built the heuristic rules to detect repeated words (Wang et al., 2024), the team did not fully train the LLM on these data. The LLM learned the several translations from the Jiji data using the few-shot prompt, whereas the baseline and SYSTRAN models adapted the output translations more directly to the target domain. Although we configured the primary results of this task based on the human evaluation, the motivation for this task was to adapt the lexical choice of MT systems to the target domain; thus, it should be noted that BLEU scores were also important metrics in our task.

Regarding the repetition rate, the trend was coincident with the human evaluation results. Specifically, the accuracy as a binary classifier (i.e., "<REP>" or not) between automatic and human evaluations was 93.4% in the baseline system, 92.1% in the SYSTRAN system, and 93.0% in the Waseda Riko system. Importantly, this metric had a certain degree of reliability independent of the success rate of style control and the degree of matching with the target domain.

8 Conclusion and Future Work

In this paper, we presented an overview of the WMT2024 Shared Task on non-repetitive translation. Particularly, the experimental results revealed the effectiveness of the LLM in controlling translation. We believe that our task will encourage further research on controlling MT systems. In the future, we will address several limitations in the current task settings. First, the test instances were limited to a comparatively short content. It would be an interesting challenge to address repetition observed in longer documents. Second, we will make both human and automatic evaluations more established. Currently, (1) evaluation relies heavily on human evaluation, and (2) the human evaluation is prone to variance. Regarding (2), specifically, although the percentage of test instances where the three translators voted for all different labels was limited, that of the test instances where the three translators voted for the same label was approximately 69%. These were partially because of (1) the difficulty of automatically detecting mistranslations and undertranslations, and (2) the difficulty of defining the correct answer for a translation output using substitution or reduction, respectively. Thus, we will develop more reliable evaluation guidelines in collaboration with translators. It would also be interesting to introduce automatic evaluation using LLMs.

Acknowledgments

We would like to thank the organizers of the WMT2024 for giving us the opportunity to present this task, the reviewers for their time and effort, and all the teams for their active participation in our task. We are also deeply grateful to Hidehiro Asaka and Takayuki Kawakami for providing the valuable data used in this research.

These research results were obtained from the commissioned research (No. 225) by National Institute of Information and Communications Technology (NICT), Japan.

References

Marko Avila and Josep Crego. 2024. Systran @ wmt24 non-repetitive translation task. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, Florida, USA. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Mervin Block. 1994. *Broadcast Newswriting: The RTNDA Reference Guide*. Bonus Books.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Jill Burstein and Magdalena Wolska. 2003. Toward evaluation of writing style: Overly repetitious word use. In *10th Conference of the European Chapter of the Association for Computational Linguistics*,

pages 35–42, Budapest, Hungary. Association for Computational Linguistics.

Rene J. Cappon. 2019. *The Associated Press Guide to News Writing*, fourth edition. Peterson's.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27, Boulder, Colorado. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. 2020. Token-level adaptive training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1035–1046, Online. Association for Computational Linguistics.

Liane Guillou. 2013. Analysing lexical consistency in translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 10–18, Sofia, Bulgaria. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. 2021. Encouraging lexical translation consistency for document-level neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3265–3277, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinglin Lyu, Junhui Li, Shimin Tao, Hao Yang, Ying Qin, and Min Zhang. 2022. Modeling consistency preference via lexical chains for document-level neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6312–6326, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710, Marseille, France. European Language Resources Association.

Toshiaki Nakazawa, Kazutaka Kinugawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Makoto Morishita, Ondřej Bojar, Akiko Eriguchi, Yusuke Oda, Chenhui Chu, and Sadao Kurohashi. 2023. Overview of the 10th workshop on Asian translation. In *Proceedings of the 10th Workshop on Asian Translation*, pages 1–28, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Anoop Kunchukuttan, Makoto Morishita, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2022. Overview of the 9th workshop on Asian translation. In *Proceedings of the 9th Workshop on Asian Translation*, pages 1–36, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45, Online. Association for Computational Linguistics.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on

Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44, Suzhou, China. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965. PMLR.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. 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.

Robert A. Papper. 2021. *Broadcast News and Writing Stylebook*, seventh edition. Routledge.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Xiao Pu, Laura Mascarell, and Andrei Popescu-Belis. 2017. Consistent translation of repeated nouns using syntactic and semantic cues. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 948–957, Valencia, Spain. Association for Computational Linguistics.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

William Strunk and E. B. White. 1999. *The Elements of Style*, fourth edition. Pearson.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Masao Utiyama and Hitoshi Isahara. 2007. A Japanese-English patent parallel corpus. In *Proceedings of Machine Translation Summit XI: Papers*, Copenhagen, Denmark.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qiao Wang, Yixuan Huang, and Zheng Yuan. 2024. Reducing redundancy in japanese-to-english translation: A multi-pipeline approach for translating repeated elements. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, Florida, USA. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Zhen Zhang, Junhui Li, Shimin Tao, and Hao Yang. 2023. Lexical translation inconsistency-aware document-level translation repair. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12492–12505, Toronto, Canada. Association for Computational Linguistics.

A Examples of Non-Repetitive Trasnlations

Table 6 shows examples of non-repetitive translations from the task website.[15]

Reduction	Ja	耐震化を済ませていない４９４団体に今後の対応を尋ねたところ、改修するのは７０団体、建て替えは２６５団体、移転が１１団体だった。
	En(trans)	When the 494 organizations that had not yet completed earthquake proofing were asked about their future measures, 70 organizations opted for retrofitting, 265 chose rebuilding, and 11 selected relocation.
	En	Of the 494 unprepared municipalities, 70 are set to carry out repairs, 265 will construct new buildings and 11 are planning relocation.
	Note	In the original English sentence, a noun ellipsis occurs, e.g., “70 municipalities” is expressed as “70.”
Reduction	Ja	開発費を参加国間で分担できるため、国産開発に比べて費用を安く抑えることが可能となる。
	En(trans)	Since development expenses can be shared among participating countries, it will be possible to keep costs lower than domestic development.
	En	It will allow the government to cut spending compared with full domestic development by sharing costs with partner countries.
	Note	“Costs” is used instead of “development costs” in the original English sentence probably because it is contextually inferable.
Reduction	Ja	同社はニューヨーク州のヨンカース工場と中西部ネブラスカ州のリンカーン工場で車両の製造や試験を行う。
	En(trans)	The company will manufacture and test vehicles at its Yonkers, New York, factory and its Lincoln, Nebraska, factory in the Midwest.
	En	Kawasaki Rail Car will build and test the subway cars at its facilities in Yonkers and in Lincoln, Nebraska.
	Note	The two nouns (“facility”) are merged into one and the noun head is shared by the two prepositional phrases. Although strictly they are not reduced, we also consider these examples to be a type of reduction.
Substitution	Ja	農作物への影響が心配されるが、農林水産省は「（首都圏などでは）積雪が長引かなかったので大きな影響はない」（園芸作物課）とみている。
	En(trans)	There are concerns about the impact on crops, but an official at the Horticultural Crops Division of the Ministry of Agriculture, Forestry and Fisheries (MAFF) said, “the snowfall (in the Tokyo metropolitan area and other regions) was not prolonged, so there will be no major impact.”
	En	Although many people are worried about the effects of harsh cold on crops, an official of Japan’s agricultural ministry predicted that there will be no significant impact, as the snow did not stay for long in areas such as the Tokyo metropolitan area.
	Note	Words with similar meaning such as synonyms and hypernyms are typically used for substitution.
Substitution	Ja	物質を構成する素粒子の振る舞いは「標準理論」で説明されるが、宇宙の全質量の４分の１を占める「暗黒物質」など説明できない部分もある。
	En(trans)	The Standard theory explains the behavior of elementary particles, which make up matter, but it cannot explain some things, such as dark matter, which makes up one quarter of the mass of the universe.
	En	The so-called Standard Model explains the behavior of elementary particles, the fundamental building blocks of matter. But the theory leaves some mysteries, such as dark matter which is thought to make up about a quarter of the mass of the universe.
	Note	Repeated words are sometimes translated in a non-literal manner.
Substitution	Ja	当時、テニス部の生徒６人とコーチがコートで練習をしており、生徒の１人がボールを拾おうとしたところ、隣のコートにパラシュート状の物があることに気付いたという。
	En(trans)	At the time, six students and the coach from the tennis club were reportedly practicing on the court when one of the students went to pick up a ball and noticed a parachute-like object on the adjacent court.
	En	At the time, the student was practicing tennis with five other students and one coach at another court next to the one where the parachute was found.
	Note	Repeated words are sometimes substituted with pronouns or pro-verbs, such as “it” and “do so.”

Table 6: Examples of non-repetitive translations from Jiji Japanese–English news articles. “Ja” and “En” indicate the original parallel sentences from the articles. “En(trans)” indicates consistent translations by humans, which are listed for comparison.

[15] https://www2.statmt.org/wmt24/non-repetitive-translation-task.html

B Detailed Statistics of the Human Evaluation Results

Table 7 shows detailed statistics of the human evaluation results.

Model			Translation Style			Total
			<NON-REP>	<REP>	<INCORRECT>	
Waseda Riko	**Adequacy (bin)**	$[5, 5]$	127	20	0	147
		$[4, 5)$	280	17	3	300
		$[3, 4)$	15	1	7	23
		$[2, 3)$	0	0	0	0
		$[1, 2)$	0	0	0	0
		Total	**422**	38	10	470
SYSTRAN	**Adequacy (bin)**	$[5, 5]$	32	45	0	77
		$[4, 5)$	71	121	5	197
		$[3, 4)$	32	66	25	123
		$[2, 3)$	16	21	29	66
		$[1, 2)$	1	0	6	7
		Total	152	253	65	470
Baseline	**Adequacy (bin)**	$[5, 5]$	66	46	0	112
		$[4, 5)$	108	53	6	167
		$[3, 4)$	43	21	41	105
		$[2, 3)$	16	9	40	65
		$[1, 2)$	3	0	18	21
		Total	236	129	105	470

Table 7: Statistics of the human evaluation results.