Fairness Testing of Machine Translation Systems

ZEYU SUN, Science & Technology on Integrated Information System Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, China
ZHENPENG CHEN, Nanyang Technological University, Singapore, Singapore
JIE ZHANG, King's College London, London, United Kingdom of Great Britain and Northern Ireland
DAN HAO, Key Laboratory of High Confidence Software Technologies (Peking University), MoE, School of Computer Science, Peking University, Beijing, China

Machine translation is integral to international communication and extensively employed in diverse human-related applications. Despite remarkable progress, fairness issues persist within current machine translation systems. In this article, we propose FairMT, an automated fairness testing approach tailored for machine translation systems. FairMT operates on the assumption that translations of semantically similar sentences, containing protected attributes from distinct demographic groups, should maintain comparable meanings. It comprises three key steps: (1) test input generation, producing inputs covering various demographic groups; (2) test oracle generation, identifying potential unfair translations based on semantic similarity measurements; and (3) regression, discerning genuine fairness issues from those caused by low-quality translation. Leveraging FairMT, we conduct an empirical study on three leading machine translation systems–Google Translate, T5, and Transformer. Our investigation uncovers up to 832, 1,984, and 2,627 unfair translations across the three systems, respectively. Intriguingly, we observe that fair translations tend to exhibit superior translation performance, challenging the conventional wisdom of a fairness-performance tradeoff prevalent in the fairness literature.

CCS Concepts: • **Software and its engineering** → **Maintaining software**; • **Computing methodologies** → **Machine learning;**

Additional Key Words and Phrases: Fairness testing, metamorphic testing, machine translation, protected attributes

ACM Reference Format:
Zeyu Sun, Zhenpeng Chen, Jie Zhang, and Dan Hao. 2024. Fairness Testing of Machine Translation Systems. *ACM Trans. Softw. Eng. Methodol.* 33, 6, Article 156 (June 2024), 27 pages. https://doi.org/10.1145/3664608

This work was supported by National Natural Science Foundation of China under Grant No. 62372005.

Authors' Contact Information: Zeyu Sun, Science & Technology on Integrated Information System Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, Beijing, China; e-mail: szy_@pku.edu.cn; Zhenpeng Chen (Corresponding author), Nanyang Technological University, Singapore, Singapore; e-mail: zhenpeng.chen@ntu.edu.sg; Jie Zhang, King's College London, London, London, United Kingdom of Great Britain and Northern Ireland; e-mail: jie.zhang@kcl.ac.uk; Dan Hao, Key Laboratory of High Confidence Software Technologies (Peking University), MoE, School of Computer Science, Peking University, Beijing, Beijing, China; e-mail: haodan@pku.edu.cn.

ACM 1049-331X/2024/06-ART156
https://doi.org/10.1145/3664608

1 INTRODUCTION

Machine learning has been highly effective in providing natural language translation systems that can handle various languages and operate in real time, with many systems capable of translating thousands of language pairs with good accuracy. Machine translation systems have a vital role in overcoming global communication barriers by providing translation services to millions of users daily. As of January 2022, Google Translate reportedly supports 133 languages, translating over 100 billion words daily [51].

However, machine translation systems have been found to have fairness issues [3]. For instance, in certain languages, Google Translate and Bing Translator have been observed to translate gender-neutral pronouns into gendered pronouns [1]. Additionally, Google Translate has produced translations that contain offensive and discriminatory content, leading to apologies from the company [2]. Table 1 provides an example of such an issue with Google Translate. When the pronoun "he" is used, the translation indicates that "he" is unlikely to "surprise" the faithful on issues such as sexuality and women's ordination. The translated term used for "surprise" in Chinese is " 惊讶 ," which aligns well with the original English sentence. However, the translation diverges when the gender pronoun changes to "she." The translated term used for "surprise" in this context is " 忠实 ," which actually means "faithful" or "loyal," not "surprise." This suggests that Google Translate may be exhibiting gender bias, as it treats the statement differently depending on the gender referred to, despite the context remaining the same. These fairness issues in existing machine translation systems are not just moral issues. They can also cause economic losses and hurt the reputation of businesses. Thus, it is crucial to automatically test the fairness issues of these systems and investigate to what extent are these systems affected by fairness issues.

Several approaches have been proposed recently to test the fairness of natural language processing [14]. However, these approaches primarily focus on classification tasks. Specifically, they aim to detect differences in the output classes by mutating fairness-related words in the input. However, these methods are not specifically designed for machine translation tasks. The primary challenge lies in the fact that its outputs are sentences—sequences of classes, as opposed to singular classes. While individual classes can be easily distinguished based on their classification results, measuring differences between two sequences of classes in sentences with different protected attributes automatically presents a challenge.

To tackle this challenge, we propose FairMT, the first framework designed to automatically test the fairness of machine translation systems based on metamorphic testing [9, 10]. FairMT is based on an assumption that translations of two semantically similar sentences, each containing protected attributes from distinct groups, should maintain comparable meanings. Based on this assumption, FairMT employs two metamorphic relations (denoted as MR1 and MR2): MR1 asserts that semantically preserving replacements of fairness-related words do not alter the sentence's meaning, while MR2 asserts the same for fairness-unrelated words, ensuring the sentence's meaning remains consistent after such substitutions. The framework contains three steps:

(1) *Test Input Generation.* This step aims to generate test inputs based on MR1 that pinpoint the fairness attributes of machine translation systems. Starting from a source sentence as the source input, we mutate fairness-related words, yielding pairs of follow-up inputs in different fairness groups. These follow-up input pairs, serving as test inputs, are then inputted into the machine translation system, generating corresponding translations.
(2) *Test Oracle Generation.* This step aims to generate test oracles corresponding to the previously generated test inputs. We use a neural-based semantic similarity metric as our test oracle. When the semantic similarity between translations of paired follow-up inputs is below a predefined threshold, we report it as a potential fairness issue. However, such issues

Table 1. An Example of a Fairness Issue on Google Translate (Collected in June 2023)

Input	Translation
A self-described obedient man of the church, he is unlikely to surprise the faithful when it comes to matters like sexuality and women's ordination.	一个自我描述的教会的听话人，在诸如性和女性命令之类的问题上，他不太可能让信徒感到惊讶。
A self-described obedient lady of the church, she is unlikely to surprise the faithful when it comes to matters like sexuality and women's ordination.	一个自我描述的教会服从女士，在诸如性和女性命令之类的事情上，她不太可能让信徒感到忠实。

may stem from low-quality translation instead of true fairness issues. This leads us to the next step.

(3) *Regression.* The step aims to discern whether these issues are truly fairness-related or merely the outcome of low-quality translation. To achieve this, we employ an additional metamorphic relation, MR2, asserting that translations of two semantically similar sentences, differing only in aspects unrelated to fairness, should preserve their analogous meanings. We use a neural-based methodology to mutating fairness-neutral segments of each sentence (the source input) in the test inputs of the potential fairness issues, and generate a set of follow-up inputs. The follow-up inputs are paired with the source input as a new test input. Then, we reprocess the test oracle step, calculating metric values for these test inputs. If the original metric score stands out as the only score beneath the threshold—indicating a discrepancy solely in the fairness-related segment—we conclusively report it as a fairness issue.

To further investigate to what extent are these systems affected by fairness issues, we conducted a detailed analysis on three state-of-the-art machine translation systems: Google Translate [22], T5 [34], and Transformer [46], using the FairMT approach. We select these models since they are the most widely used in academia or industry. Our analysis focused on translations between English and Chinese, the two most widely spoken languages with over one billion speakers worldwide [51]. The experimental results show that all three translators have fairness issues, where FairMT detected up to 832, 1,984, and 2,627 fairness issues on Google Translate, T5, and Transformer, respectively. The further human evaluation shows the validity of the detected issues. We also found that there is a consistent positive relationship between BLEU scores and the fairness similarity metric as indicated by an average estimate of 1.95. Translations without fairness issues tend to be of higher quality, which is contrary to the widely recognized fairness-performance tradeoff in the fairness literature [7, 12–15, 27, 50]. For Google Translate/T5/Transformer, legal translations[1] achieve averagely 20%/21%/30% higher BLEU score than those with detected fairness issues on two fairness groups.

To sum up, this article makes the following contributions:

- *Introduction of FairMT Framework.* This article presents a metamorphic-testing-based framework specifically designed to test the fairness in machine translation systems. This fills a gap in existing literature, where existing fairness testing methods cannot be used on machine translation tasks.
- *Empirical Evaluation on Leading Systems.* This article conducts an in-depth empirical analysis of three state-of-the-art machine translation systems, providing new insights into the prevalence of fairness issues even in high-performing translation systems.
- *Available Data.* The code and data used in our experiments are available at our homepage https://github.com/zys-szy/FairMT.

[1] Legal translations refer to translations where no fairness issues have been detected.

The structure of this article is outlined as follows: Section 2 shows relevant work to fairness testing and machine translation testing. Section 3 provides an in-depth description of FairMT, our proposed methodology. To assess the fairness of various machine translation systems, we conducted a series of experiments utilizing FairMT, detailed in Section 4. The results of these experiments are presented in Section 5. Finally, this article discusses potential threats to the validity of our study in Section 6 and concludes with key findings and implications in Section 7.

2 RELATED WORK

This article focuses on fairness testing of machine translation systems, which lies in the intersection of two significant software testing areas: machine translation testing and fairness testing.

2.1 Fairness Testing

Fairness testing research primarily focuses on the automated generation of test inputs capable of exposing discrimination and unfairness issues in software systems [11]. This field has garnered increasing interest within the SE community, particularly following the groundbreaking work by Galhotra et al. [21] at ESEC/FSE 2017, which introduces the first fairness testing approach for ML software in SE and received the Distinguished Paper Award. This approach employs a random test input generation technique to identify discriminatory instances from the input space.

In addition to the random generation approach, recent research in fairness testing techniques has embraced more advanced methods, notably the adoption of search-based input generation approaches [4, 20, 29, 43, 44, 54–56, 58]. For instance, Monjezi et al. [29] introduce an information-theoretic search-based approach, which employs gradient-guided clustering to explore the input space and seek instances that exhibit the most pronounced individual discrimination quantitatively; Aggarwal et al. [4] use a decision tree to approximate the decision-making process of ML software and applied symbolic execution to traverse various paths within the decision tree, aimed at identifying inputs that lead to discriminatory outcomes.

Despite extensive efforts in fairness testing, the majority of such testing has primarily focused on tabular data, while the realm of fairness testing for text-based systems remains relatively nascent [11]. Recently, researchers are beginning to address this gap. For example, Asyrofi et al. [5] propose a test input generation approach to uncover fairness bugs in sentiment analysis systems. It uses NLP techniques to identify words associated with demographic characteristics and replaces them with different values to form test input. Soremekun et al. [39] propose a grammar-based fairness testing approach. They leverage context-free grammars to generate equivalent test inputs and apply metamorphic relations to evaluate equivalent test inputs for software fairness. Wan et al. [47] propose a test input generation approach aimed at fairness testing of conversational systems. They design templates and rules for generating questions for conversational systems. However, these methods cannot be applied to machine translation systems, since they face challenges in measuring differences between two sequences of classes in sentences with different protected attributes automatically.

2.2 Machine Translation Testing

Existing machine translation testing approaches focus on testing different properties to test in machine translation systems. These approaches can be divided into testing on multiple translation systems and testing on a single translation system.

Testing on Multiple Translation Systems. Pesu et al. [32] introduce a cross-reference approach. They examine the consistency between direct translation (from a source language to a target language) and indirect translation (translating from the source language to an intermediate one and subsequently from the intermediate to the target language) of an identical sentence to determine if they yield consistent outcomes. Cao et al. [8] assess the similarity in translation outcomes when different translation systems are employed. They investigate if these varied systems produce similar translation results.

Testing on a Single Translation System. Heigold et al. [26], Belinkov and Bisk [6], and Zhao et al. [57] concentrate on evaluating the robustness of machine translation systems. Specifically, they investigate the sensitivity of these translators to slight perturbations, such as minor errors, typos, or noise in the input sentences.

Sun and Zhou [40] replace one human name with another in phrases preceding words such as "like" or "hate". They expect that the translations, before and after this word replacement, should remain largely consistent. Similarly, TransRepair [41] replaces one word with another that shares semantic similarities. The expectation here is that the translated output for parts of the sentence that remained unchanged is still consistent closely to the original translation. Expanding upon this idea, CAT [42] introduces a context-aware replacement strategy termed "isotopic replacement" for word replacement. Its output expectations mirrored those of TransRepair. Wang et al. [48] focus on the word sense disambiguation ability of machine translation systems and aim to test the exact sense of polysemes (i.e., words with multiple senses) in the given context. Purity [25] dissects original sentences into distinct phrases, examining whether standalone translations of these phrases diverged from their translations within the full original sentence. Delving into the structure of sentences, He et al. [24] choose to replace a word in a given sentence with another word. They expect that translations of both the original and the replaced sentences should have structural similarities. Gupta et al. [23] focus on a different aspect: detecting translation inaccuracies by replacing a word with another of entirely different semantics. They expect that such a change should lead to different translations.

Many approaches to machine translation testing have been put forward over time. However, a conspicuous gap exists: none specifically focuses on the fairness aspect of machine translation systems. In light of this, we introduce FairMT. To the best of our knowledge, we are the first to automatically test the fairness issue of machine translation systems.

2.3 Fairness Testing for Machine Translation Systems

There is no automated fairness testing method designed specifically for machine translation systems. Current fairness testing for these systems primarily relies on manually crafted templates. For instance, Wang et al. [49] manually create 30 templates with placeholders for person names, utilizing names from different genders to assess the machine translation system's proficiency in gender estimation during translation. Similarly, Prates et al. [33] develop templates such as "He/She is an [job]" (where [job] represents a job position of interest). They compile a list of relevant job positions to populate these templates and generate test inputs. Then, they translate these generated inputs into English and collected statistical data on the frequency of male, female, and gender-neutral pronouns in the translated output, enabling the measurement of machine translation bias. To address this research gap, this paper introduces an automated fairness testing approach tailored specifically to machine translation systems.

3 APPROACH

We use the metamorphic testing and assume that translations of two semantically similar sentences, even when they possess protected attributes from different groups, should retain similar meanings. Based on this assumption, we propose FairMT, a framework to test the fairness of machine translation systems. FairMT employs two metamorphic relations (denoted as MR1 and MR2): MR1 asserts that semantically preserving replacements of fairness-related words do not alter the sentence's meaning, while MR2 asserts the same for fairness-unrelated words, ensuring the

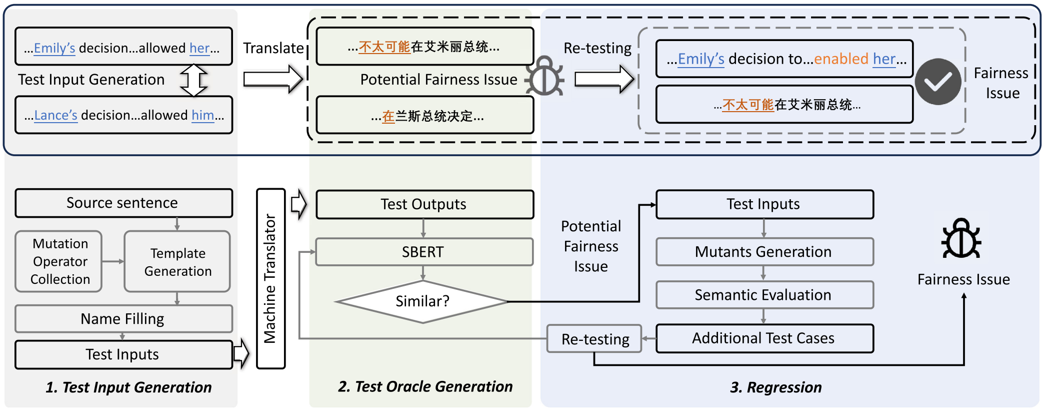

Fig. 1. The overview of FairMT.

sentence's meaning remains consistent after such substitutions. In this section, we first define the fairness issue and then introduce FairMT.

3.1 Fairness Definition

In this paper, we focus on the individual fairness definition that is widely adopted in the fairness testing literature [4, 5, 20, 21, 43, 44, 55]. It involves altering the sensitive attribute(s), and assessing if this alteration affects the output. Fairness issue occurs when two similar individuals, which differ only in their protected attributes, are treated unequally by a model [11].

Let X represent the input, with $A \subset X$ denoting the words corresponding to protected attributes within the sentence, and $N \subset X$ representing the remaining words. For an input sentence $x = \{x_1, x_2, \ldots, x_n\}$, a fairness issue in a machine translation system f arises when there exists another similar sentence x' that satisfies the following conditions:

$$f(x) \neq f(x') \quad s.t. \exists x_p \in A, x_p \neq x'_p; \forall x_q \in N, x_q = x'_q. \tag{1}$$

In our context, $f(x) \neq f(x')$ denotes the translations of these two inputs are not equal in semantics; $x_p \neq x'_p$ denotes the words are in different demographic groups; $x_q = x'_q$ denotes the words are equal in semantics.

3.2 Overview

Figure 1 shows an overview of FairMT, which contains the following three steps.

(1) *Test Input Generation.* This step aims to generate mutated sentences that serve as test inputs for detecting the fairness of machine translation systems. The generation follows the MR1, within each source sentence as source input, FairMT mutates fairness-related words by substituting them with a pair of the fairness-related words in different groups (e.g., a pair of a male name and a female name) via a template-based approach. The generated follow-up inputs are paired as the test inputs for further computation. Details are presented in Section 3.3.
(2) *Test Oracle Generation.* This step aims to generate test oracles for the previously generated test inputs. These test oracles are instrumental in detecting fairness issues within the system. Initially, the generated test inputs are inputted into the translation system, yielding output translations. Drawing inspiration from metamorphic testing, we identify candidate fairness issues based on the semantic similarity score between the translations of two

sentences within the test inputs, utilizing a neural-based semantic evaluation metric. If the score falls below a predefined threshold, the test case–comprising input sentences and output translations–is reported as the potential fairness issue. Details are presented in Section 3.4.

(3) *Regression.* This step aims to check whether the reported potential fairness issues are truly fairness issues rather than the issues caused by the low-quality translation. We employ an additional metamorphic relation, MR2, asserting that translations of two semantically similar sentences, differing only in aspects unrelated to fairness, should preserve their analogous meanings. If a potential fairness issue arises that also does not adhere to this relation, it can be indicative of a translation of inferior quality. For each potential fairness issue detected in the previous step, we re-generate test inputs by mutating words that are not fairness-related words within each sentence in the test inputs of the issue. Then, we repeat the test oracle generation process for the second testing process and get the scores. If the original score of the potential fairness issue is the only score that is below the predefined threshold, this issue is reported as the final fairness issue. Details are presented in Section 3.5.

3.3 Test Input Generation

In this article, we focus on two protected attributes: gender and country-of-origin. Given a source sentence S as the source input, we aim to mutate these two protected attributes present within it. To achieve this, we use the metamorphic relation MR1, which asserts that semantically preserving replacements of fairness-related words do not alter the sentence's meaning. Then, we use a template-based test input generation methodology called BiasFinder [5]. It generates test inputs by recognizing the names within the source sentence and mutate them into the names in different groups, thereby introducing mutations that reflect different genders and countries.

The test input generation process contains the following steps:

(1) *Mutation Operator Collection.* Focusing on two protected attributes, namely gender and country-of-origin, we first build a set of mutation operators. These are designed to substitute a name in the source sentence with two distinct names, each representing a different fairness group. For instance, one could be a female name while the other is a male name, or they may be names from two different countries.

To achieve this, we use the name set provided by BiasFinder [5]. This set contains 30 female and 30 male names from the USA for gender-based fairness. Additionally, it also includes 25 female and 25 male names from different countries for country-based fairness. Consequently, using these sets, we have (1) 900 mutation operators for pairing a female name with a male name, (2) 625 mutation operators for pairing two female names from different countries, and (3) 625 mutation operators for pairing two male names from different countries.

(2) *Template Generation.* After gathering a set of mutation operators for name replacement, our next step is to pinpoint relevant words within the provided source input S. We then generate a template by mutating these identified words with placeholders, setting the stage for the eventual insertion of protected attributes.

To detect these fairness-related words, we employ a named entity recognition method [30] to automatically identify names within the sentence S. Following this, we use a coreference resolution technique [38] to dissect the sentence and spot all expressions that link to the same name.[2] For instance, in the sentence "Alice has a cute dog. She loves it.", the words "Alice" and "She" are connected since both refer to "Alice". Using these identified expressions, we then replace these words with placeholders, forming a template T. For instance, we can get a

[2] Following BiasFinder, we consider the gender-related terms listed in https://github.com/zys-szy/FairMT/blob/main/wordsInTestGeneration.txt

template "<name> has a cute dog. <She> loves it." from the former instance, where "<name>" denotes the placeholder that can only fill a name and "<She>" denotes the placeholder that can only fill the words in the same type, namely "She" or "He". Notably, if multiple names appear in S, only the first one is replaced.

(3) *Name Filling.* Building upon the template established in the prior step, our objective in this phase is to replace the placeholders within the template with names sourced during the mutation operator collection.

For each template, we fill a name to the placeholder. Concurrently, any correlated placeholders referencing this name also undergo a similar substitution. For example, given the template "<name> has a cute dog. <She> loves it.", we can replace the placeholders with "Bob" and "He" respectively, resulting in the sentence "Bob has a cute dog. He loves it.". Our primary goal during this phase is to fill the placeholders using pairs of names representative of distinct fairness groups (i.e., different genders or different country-of-origin), to yield pairs of follow-up inputs.

For gender, we employ the 900 collected female-male name pairs as fill-ins. If the source sentence lacks a name, we only substitute placeholders for gender-related terms. For example, in the template "<She> loves it," the placeholder <She> would be replaced with either "He" or "She". For country-of-origin, we employ 625 pairs of female and male names from different countries. In instances where the source sentence does not contain a name, we leave the placeholders untouched.

We proceed to fill the names into the template repeatedly for N iterations (in this article, N is set at 10). In particular, for country-of-origin, we repeat this process for $N/2$ iterations on female names and $N/2$ iterations on male names, respectively. During each iteration, we randomly select a pair of names from our collection and use them to fill the respective placeholders. This process outputs two distinct follow-up inputs that together form a test input. After cycling through this procedure N times, we are left with a set of test inputs, T, where $|T| = N$.

3.4 Test Oracle Generation

After obtaining these test inputs T, the next step aims to gather the test outputs for these inputs, along with their corresponding test oracles. This allows us to check whether there are fairness issues in each test case.

For each sentence S_i within the test inputs $T_i = \{t_1, t_2\}$, we first feed it into the machine translation system (e.g., Google Translation and Bing Translator) to generate output translations $O_i = \{o_1, o_2\}$.

To check whether there are fairness issues in these test cases, we note that the test input sentences, differing only slightly in terms of fairness-related words, should essentially convey similar semantics. This similarity should ideally extend to their translations as well. Thus, assessing the semantic similarities of the translations serves as a reliable method to uncover potential fairness issues. In this article, we employ the widely used SBERT [35], an algorithm specifically devised for deriving semantically meaningful sentence embeddings. Comparisons between these embeddings can be made using cosine similarity as a measure of semantic similarities.

For every pair of translated outputs, denoted as $O_i = \{o_1, o_2\}$, we input these translations into SBERT, obtaining their corresponding representations as real-valued vectors, $\{\boldsymbol{o}_1, \boldsymbol{o}_2\}$. We then compute semantic similarity based on these representations via cosine similarity

$$\mathrm{Sim}(\boldsymbol{o}_1, \boldsymbol{o}_2) = \frac{\boldsymbol{o}_1 \boldsymbol{o}_2}{|\boldsymbol{o}_1||\boldsymbol{o}_2|}. \tag{2}$$

If the similarity $\mathrm{Sim}(\boldsymbol{o}_1, \boldsymbol{o}_2)$ for a test case is below a pre-defined threshold D, the test case is reported as a "potential" fairness issue. We emphasize "potential" since, at this juncture, it remains undetermined if the reported issue arises from low-quality translation or represents a genuine fairness issue. We compute the similarity of all the test cases and report a set of potential fairness issues I.

3.5 Regression

This step is to discern if previously reported potential fairness issues are genuine fairness issues or merely caused by low-quality translation. This implies that the semantic differences between the translations arise solely from replacing one fairness-related word for another from a different group, and not from replacing words that are unrelated to fairness. For instance, after detecting a potential fairness issue in the translation differences between "The male engineer fixed the problem" and "The female engineer fixed the problem," we apply regression testing by changing a non-fairness-related word, altering "fixed" to "solved" in both sentences. This change, "The male engineer solved the problem" tests the translation's quality without altering the fairness-related words. If the translation between the changed one and the original one still contains the same issue, it indicates that the observed discrepancy indeed stems from low-quality translation, rather than an issue related to the fairness-related words. This finding suggests that the initial potential fairness issue is more likely due to inconsistencies in translation quality across different verb choices or sentence constructions, rather than discriminatory treatment based on the gender implied by the fairness-related terms.

To realize this idea, we employ an additional metamorphic relation, MR2, asserting that translations of two semantically similar sentences, differing only in aspects unrelated to fairness, should preserve their analogous meanings. If a potential fairness concern arise that also does not adhere to this relation, it could be indicative of a translation of inferior quality.

For testing, we employ regression testing principles but incorporate additional test cases. We re-test the machine translation system for these new cases. The test inputs of these test cases are generated by mutating the fairness-unrelated words within a test case of the reported potential fairness issues. For the test oracles, We reuse the test oracles from the prior subsection "Test Oracle Generation" for these test cases. Following re-testing, if the original test case stands out as the sole issue among all other test cases, we classify it as a fairness issue. The details are introduced in the following section:

3.5.1 Test Inputs of the Additional Test Cases. For test inputs of the additional test cases, we first need to ensure that replacing the fairness-unrelated words will not lead to a change in the semantics.

For each sentence t in a potential fairness issue, we first take it as source input and find all fairness-related words including named entities, personal pronouns, races, ages, and occupations. For named entities (e.g., personal names, cities, and countries), we use the name entity recognition method [37] to recognize all named entities within the sentence. For others, we use a set of collected in total 393 words to recognize other fairness-related words within the sentence via string match. For the ages, we additionally use a template of "<placeholder>-year-old", where the "<placeholder>" can be replaced by a number. Then, we filter these words and in t get remaining fairness-unrelated words $W_{(\mathrm{uf})}$.

After getting these words, the next step is to replace these words with others to test whether the semantic differences between the translations arise solely from replacing one fairness-related word for another from a different one. Thus, the replacement should consider the context of the sentence and keep the semantics of t unchanged.

To achieve this, we use a context-aware word replacement approach (CAT) [42] for word replacement. Given a sentence, CAT uses BERT [59] to mutate each word within the sentence and generates a set of mutant sentences (Mutants Generation). For each mutant, it then computes the semantic similarity between the mutated word and the original word within the sentence (Semantic Evaluation). It leaves only the mutants with the high similarity as the follow-up inputs.

Mutant Generation. In detail, for a sentence t, we first tokenize it into a sequence of tokens $w_1, w_2, \cdots, w_K$, where K denotes the length of the sentence. For each fairness-unrelated word $w_i \in W_{(\mathrm{uf})}$, we in turn mask it as a special token "[MASK]" (only one word being masked each time) and feed the masked sentence to the BERT. The outputs of BERT are a sequence of word vectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_K$, which denotes the representations of the input words. For the masked token $w_{(\mathrm{mask})}$, we use the pre-trained classifier, which is pre-trained from **MLM (Masked Language Model**; i.e., predicting a suitable word to replace the "[MASK]") task of BERT, to predict a set of suitable words. Each word contains a predictive probability. The sum of the probabilities of all the candidate words is 1.0. We then use a probability filter to discard the words with low predictive probability (in this article, we discard the words with a probability below 0.01). The remaining words are used to fill the "[MASK]" and get mutants. After masking all fairness-unrelated words, we get a set of mutants M.

Semantic Evaluation. For each mutant $m \in M$, we feed both the mutant and the source sentence t to BERT, each of them extracting a sequence of word vectors. From these two sequences, we can select both the word vector of the mutated word in the mutant $\boldsymbol{w}_m$ and the word vector of the original word in the source sentence $\boldsymbol{w}_s$. These vectors represent the semantics of the words. To ensure that the word replacement will not change the semantics of the sentence, CAT computes the semantic similarity between these vectors via cosine similarity. The computing is similar to Equation (2) but with different inputs. If the similarity is beyond the threshold (we set the threshold at 0.7), we leave the mutant as the follow-up inputs and pair it with the source input as a new test input. We leave at most five mutants for the source sentence. Since each potential fairness issue is uncovered by an input pair with two sentences, we have up to 10 additional test inputs for further re-testing. After the computation for all mutants, we have a set of new test inputs T_n for regression.

3.5.2 Re-testing. Re-testing aims to re-test the machine translation using the newly generated test inputs T_n of each potential fairness issue and decide whether the reported potential fairness issue is truly a fairness issue. For these test inputs, we repeat the entire process of test oracle generation in Section 3.4 and get a set of the output semantic similarities. For each similarity, we compared it with a predefined threshold D. If there is any similarity beyond the threshold, we discard the corresponding reported potential fairness issue. Finally, after the computation for all potential fairness issues, we report the remaining issues as fairness issues.

4 EXPERIMENTAL SETUP

In this section, we introduce the procedure we use FairMT to evaluate several machine translation systems.

4.1 Research Questions

To test the fairness of existing machine translation systems, we use FairMT on them and investigate three research questions.

- *RQ1: How effective is FairMT in generating test inputs?*
- *RQ2: To what extent are current machine translation systems affected by fairness issues?*

— *RQ3: What are the relationships between performance and fairness?*

To answer these research questions, we employ FairMT to detect gender-related/country-of-origin-related fairness issues in existing machine translation systems. We select these two attributes because gender and country of origin are among the most common dimensions where biases are observed in datasets and AI models [5].

As introduced in the Section 3.4, FairMT needs a pre-defined similarity threshold D. In this section, we conduct experiments on four different similarity thresholds and present the number of detected fairness issues across them. Then, to make sure these issues are valid, we further sample some of them and manually check the validity of them. The results show that all selected machine translation systems contain fairness issues and FairMT is effective in detecting them. Further, we delve deeply to comprehensively analyze the detected fairness issues.

Upon evaluating the fairness issues, we find that as the proficiency of a translation system increases, the number of fairness issues it presents appears to decrease. This is contrary to a prevailing belief in the field. Typically, it is assumed that there exist fairness-performance trade-offs in ML. Thus, in RQ3, we aim to understand this observation and focus on the relationships between translation ability and fairness.

4.2 Dataset

We use the News Commentary dataset [52], a widely adopted large-scale benchmark for machine translation testing [41, 42]. This dataset comprises parallel corpora collected from news commentary sources. It features articles and commentary from international news, providing rich, real-world text samples. The languages included are Arabic, Chinese, Czech, Dutch, English, French, German, Hindi, Indonesian, Italian, Japanese, Kazakh, Portuguese, Russian, and Spanish. Consistent with previous studies [41, 42], we focus on the English-Chinese translation subset, because they rank as the top two most widely spoken, each boasting over one billion speakers globally [18]. The training set contains 361,456 parallel sentences. Due to the computation cost, we randomly selected 5,406 (1.5%) English-Chinese parallel sentences from the News Commentary dataset for our experiments. To the best of our knowledge, the size of the data we use is the largest in the machine translation testing literature [23–25, 41, 42].

4.3 Machine Translation Systems

In this article, we conducted experiments on three state-of-the-art machine translation systems: (1) *Transformer* [46]: The most widely used machine translation system in academic research; (2) *T5* [34]: A widely used large language model based on Transformer; and (3) *Google Translate* [22]: The most widely used end-to-end machine translation systems developed by Google.

Transformer. Transformer is an attention-based sequence-to-sequence model widely used in natural language processing tasks. It shows good performance across diverse domains including machine translation [46], chatting [31], and more [17]. For the implementation, we use the transformer model in Tensor2Tensor library [45] with the same parameters for the Transformer model as described in [41]. The model is trained based on three datasets: the CWMT dataset [16] with 7,086,820 parallel sentences, the UN dataset [60] with 15,886,041 parallel sentences, and the News Commentary dataset [52] with 252,777 parallel sentences. The validation set is from the News Commentary dataset with 2,002 parallel sentences.

T5. **T5 (Text-to-Text Transfer Transformer)** is a pre-trained model based on transformer. It shows good performance in generation tasks [34]. For the implementation, we use the pre-trained T5 in the transformers library [53]. This model is fine-tuned on the News Commentary dataset [52]

Table 2. Number of Generated Test Inputs (RQ1)

Group	#Sentences	#Test Inputs	#Test Inputs/sentence	Accuracy
Gender	5,406	36,514	6.75	99%
Country of Origin		34,120	6.31	100%

with 252,777 parallel sentences. The validation set is from the News Commentary dataset with 2,002 parallel sentences.

Google Translate. Google Translate. Google Translate is a machine translation product of Google. It stands out due to it is widely used in the realm of translation. As of 2022, Google Translate offers support for 133 languages at various levels. By April 2016, it boasted over 500 million users, translating more than 100 billion words daily [51].

4.4 Implementation Settings

The BERT model used in FairMT is based on a public pre-trained BERT in the transformers library [53]. BERT contains 24 layers. The hidden size is set to 1,024 and the multi-head attention contains 16 heads. It was trained on 16 TPU chips for one million steps with a batch size of 256 [59]. For Google Translate, we collected the translations from June to September 2023.

We conduct experiments on Ubuntu 16.04 with 256GB RAM and four Intel E5-2620 v4 CPUs. The neural networks used in our experiment including BERT, transformer, and T5 are all trained and inference on a single Nvidia Titan RTX. In particular, Transformer and T5 are trained on the entire training set, which contains 361,456 parallel sentences

5 RESULTS

In this section, we introduce the results of testing the fairness issues on different machine translation systems.

5.1 Effectiveness of Test Input Generation (RQ1)

To answer this question, we use the News Commentary dataset, encompassing 5,406 sentences in total. Each sentence from this dataset serves as a source sentence. Leveraging FairMT, we focus on generating test inputs used for both gender-based and country of origin-based fairness issues detection.

Our evaluation process consists of two aspects: (1) Quantity: First, we examine the number of test inputs generated by FairMT across these machine translation systems. (2) Validity: To ensure the validity of the generated test inputs, we manually sample and verify a subset of these cases. We assess the validity based on the principle that mutations should not compromise the underlying meaning of the sentences.

5.1.1 Quantity. The results are shown in Table 2. As shown, FairMT successfully mutates the original set of 5,406 sentences into a substantial number of test inputs, totaling 36,514 and 34,120 for gender and country of origin groups, respectively. Consequently, this mutation yields an average of 6.75 test inputs per original sentence for the gender group and 6.31 for the country of origin group. For example, FairMT replaces "Leroy" and "his" in "But make no mistake: both parties are implicated. There is already talk that Leroy will raise $1 billion or more for his re-election campaign." to "Carolyn" and "her".

5.1.2 Validity. To assess validity of these generated test inputs, we further randomly selected 100 of these test inputs and manually evaluated their correctness on both groups, respectively. We determined the correctness of each test input based on the criteria: if the mutations within the test inputs preserved the semantics and the protected attributes of the unchanged parts, we labeled the test inputs as correct. Conversely, if these mutations altered or compromised these aspects, the inputs were deemed incorrect. Following this evaluation, for gender-based test inputs, we found that the generation accuracy stood at 99%, with only a single anomaly identified. Specifically, FairMT erroneously replaced “She” with “He” in the sentence “She has been called the Queen of Elections.” This modification was deemed inappropriate, as the term “He” does not align with the title “Queen of Elections.” For country of origin-based test inputs, our evaluation revealed a 100% accuracy rate, attributable to the relative ease of substituting country names within sentences. This result shows the good performance of the test input generation step in FairMT.

In the next section, we use these generated test inputs to examine the fairness of machine translation systems.

> **Answer to RQ1:** FairMT successfully generates 36,514 test inputs for the gender group and 34,120 test inputs for the country of origin groups based on the original dataset of 5,406 sentences. This equals an average of 6.75 and 6.31 test inputs per original sentence for each group. Manual evaluation indicates that 99% and 100% of the generated test inputs are accurate for the two groups, respectively.

5.2 The Fairness Issues of Machine Translation Systems (RQ2)

To answer this question, leveraging the former generated test inputs, we use steps 2 (test oracle generation) and 3 (regression) of FairMT to detect the gender-based and country of origin-based fairness issues within three widely used machine translation systems: Google Translate, Transformer, and T5.

Our evaluation process consists of multiple aspects: (1) *Quantity*: First, we examine the number of fairness issues detected by FairMT across these machine translation systems. We also scrutinize various similarity thresholds to better understand the scope and extent of potential fairness issues; (2) *Validity*: To ensure the validity of the detected issues, we manually sample and verify a subset of these cases. We assess the validity based on the semantic differences in the translations generated by the machine translation systems; and (3) *Analysis*: We delve into a qualitative analysis of these fairness issues. By comparing translations and studying individual cases, we gain a comprehensive understanding of these fairness problems.

5.2.1 Quantity. We first investigate the contributions of different steps in FairMT, along with the quantity of it across varied similarity thresholds.

In the *test oracle generation* step, we feed these generated test inputs to Google Translate, Transformer, and T5. After translation, we generate test oracles to discern if the translations have potential fairness issues. In our experiment, we configure four distinct thresholds, denoted as D, to detect such potential issues. The results are shown in Table 3, labeled as #Pot. FI (the number of potential fairness issues). As shown, for the gender group, FairMT detects 46/107/283/964, 274/521/1,019/2,320, and 432/751/1,296/2,287 potential fairness issues on Google Translate, T5, and Transformer with the threshold 0.70/0.75/0.80/0.85, respectively. For the country of origin group, FairMT detected 23/46/107/307, 246/409/731/1,308, and 449/718/1,161/1,817 potential fairness issues.

We further feed these potential fairness issues to the third step (i.e., *regression*). In this step, FairMT first generates additional test cases for each potential fairness issue. FairMT generates additional test cases by substituting words that are not related to fairness with semantically similar counterparts. Specifically, FairMT avoids replacing fairness-related words such as “he” to prevent

Table 3. Number of Reported Fairness Issues with Different Thresholds between 0.70 and 0.85 (RQ2)

		Threshold											
		0.70			0.75			0.80			0.85		
	Machine Translation	#Pot. FI	#Reg.	#FI	#Pot. FI	#Reg.	#FI	#Pot. FI	#Reg.	#FI	#Pot. FI	#Reg.	#FI
Gender	Google Translate	46	265	39	107	635	81	283	1,615	211	964	5,170	659
	T5	274	1,575	166	521	2,910	316	1,019	5,685	622	2,320	12,460	1,345
	Transformer	432	2,325	338	751	4,015	560	1,296	7,005	913	2,287	12,095	1,469
Country	Google Translate	23	115	11	46	235	26	107	590	64	307	1,610	173
	T5	246	1,350	150	409	2,225	237	731	3,910	404	1,308	6,850	639
	Transformer	449	2,365	354	718	3,800	543	1,161	6,135	827	1,817	9,520	1,158

The first column categorizes the data by fairness group including gender and country of origin. The second column lists the evaluated machine translation systems. The subsequent columns detail three key metrics: the Number of Potential Fairness Issues (#Pot. FI), the Number of Additional Test Cases (#Reg.), and the Number of Final Fairness Issues (#FI).

introducing fairness issues. However, it can substitute other words, like "This," with any suitable terms, including "he," provided that such replacements do not alter the sentence's semantics or affect the fairness context. The number of test cases is shown in Table 3, labeled as #Reg. Overall, FairMT generates an average of 5.49/5.28, 5.47/5.32, and 5.34/5.26 additional test cases for each potential fairness issue detected in Google Translate, T5, and Transformer on the group of gender/country of origin. We conducted manual checks for the correctness of these additional test cases by randomly selecting 100 cases for each machine translation system and each fairness group. Our evaluation criteria were twofold: First, the semantics of the sentences must remain intact; second, the alterations should not adversely affect the fairness-related parts. Our findings revealed that the accuracy rates for these test cases were 97%/93%, 95%/91%, and 96%/97% for Transformer, T5, and Google Translate in the group of gender/country of origin, respectively. Such results show the good effectiveness of FairMT. Further analysis revealed that inaccuracies often arose from FairMT improperly modifying fairness-related parts. For instance, in the sentence "He was the same Budivid who in 1996 ...," FairMT erroneously replaced "He" with "This."

We then use all additional test cases for re-testing. The number of detected fairness issues is shown in Table 3, labeled as #FI. We observe that FairMT successfully detects fairness issues on all three machine translation systems and all thresholds. FairMT effectively detects up to 832, 1,984, and 2,627 fairness issues across the three systems, respectively. For the gender group, FairMT detects 39/81/211/695, 166/316/622/1,345, and 338/560/913/1,469 fairness issues on Google Translate, T5, and Transformer with the threshold 0.70/0.75/0.80/0.85, respectively. For the country of origin group, FairMT detects 11/26/64/173, 150/237/404/639, and 353/543/827/1,158 fairness issues in Google Translate, T5, and Transformer at thresholds 0.70/0.75/0.80/0.85, respectively. This result shows that no matter how powerful a machine translation system is, there are still fairness issues within it. The number of detected fairness issues is an average of 70.71%/56.73%, 59.24%/53.08%, and 68.82%/69.53% of the number of the detected potential fairness issues on Google Translate, T5, and Transformer in the group of gender/country of origin. This shows that FairMT filters some issues caused by the low-quality translation rather than fairness.

We further show two examples of the low-quality translation. Google Translate correctly translates "Nor is Mildred promise to enact immediate cuts in federal discretionary spending by an additional 5% likely to boost job growth..." to " 米尔德雷德（Mildred）也没有承诺在联邦酌情支出中立即削减5％的削减... ". When we replace to female name "Mildred" to the male name "Zachary", the translation " 米尔德雷德（Mildred）也没有承诺 ... ", which means "Nor is Mildred promise ..." changes to " 扎卡里（Zachary）也承诺 ... ", which means "Zachary promises

...”. The similarity score between these two translations is 0.67. It seems like a fairness issue. However, FairMT finds that this phenomenon also exists when we change the fairness-unrelated words. FairMT replaces the word “additional” with “extra”, which does not change the semantics of the input sentence. The translation is also changed to “ 米尔德雷德（Mildred）也承诺 ... ”, which also means “Mildred promises ...”. The similarity score between these two translations is slightly below 0.70. Thus, FairMT filters it, since this may caused solely by the low-quality translation. For Transformer, it translates “... this embarrassment ... Lance ... reveals a level of incompetence beyond anything suspected so far.” into Chinese misleadingly suggests, “ ... 兰斯 ... 显示出了远远超出任何怀疑的能力。 ”, which means “Lance demonstrates a capability far beyond any doubt.” However, upon replacing “Lance” with a female name “Julie” (this is done by the first two steps of FairMT), the translation shifts to “ ... 朱莉 ... 表现出一种无能。 ”, which reverts to “... Julie ... exhibits a kind of incompetence.” This change, resulting in a significant alteration in narrative tone based on the gender of the name, highlights critical fairness concerns. It implies that the translation model may process gendered contexts in a biased manner, potentially reinforcing stereotypes or biases. FairMT effectively detects it as a potential fairness issue, applying a threshold of 0.80 for the semantic similarity score. The issue is flagged with a score of 0.79, which is below the set threshold. Nonetheless, the regression step of FairMT reveals an intriguing finding: altering the source sentence from “... this embarrassment ... Lance ... reveals a level of incompetence beyond anything suspected so far.” to “... such embarrassment ... Lance ... reveals a level of incompetence beyond anything suspected so far.” (changing “this” to “such”) also alters the translation to “ ... 兰斯 ... 显示出一种无能的程度。”, mirroring the meaning attributed to the modification from “Lance” to “Julie”. This indicates that the observed discrepancy is attributed to translation quality rather than inherent fairness bias. Consequently, FairMT gets a semantic similarity score of 0.79, which falls below the threshold of 0.80, and excludes it from the list of fairness issues.

We also observe a negative correlation between the number of detected fairness issues and the machine translation ability. Specifically, Google Translate, being the most effective machine translation system among the three we selected, exhibits the fewest fairness issues. When compared to T5 and Transformer, Google Translate’s performance is better. For the gender group, across the four thresholds, Google Translate contains 59.58% fewer fairness issues than T5 (which is a pre-trained version of Transformer and is acknowledged to outperform the original model) and 69.82% fewer than the Transformer. For the country of origin, across the four similarity thresholds we tested, Google Translate demonstrated significantly fewer fairness issues: 80.84% fewer than T5 and 90.49% fewer than Transformer.

5.2.2 Validity. In this section, we evaluate the validity of the fairness issues detected by FairMT across three major machine translation systems–Google Translate, T5, and Transformer. We examine these systems at four different similarity thresholds: 0.70, 0.75, 0.80, and 0.85. To perform this evaluation, we randomly select 50 fairness issues for each system at each threshold. In particular, if the number of fairness issues is less than 50, we select all of them. We manually review them for validity based on two key metrics:

- *#FI (Number of Fairness Issues)*: This denotes the number of test cases where fairness issues are detected. We assess the validity of these fairness issues by examining the semantic differences between the translations of paired sentences. If the semantic content between the translations diverges, we classify it as a fairness issue.
- *#DIS (Number of Discriminations)*: Discriminations denote a subset of fairness issues. A case is reported as discrimination if the translation appears to unfairly favor or disfavor a specific gender.

Table 4. The Human Evaluation Results of the Reported Fairness Issues with Different Thresholds between 0.70 and 0.85 (RQ2)

		Threshold							
		0.70		0.75		0.80		0.85	
Machine Translation		#FI	#DIS.	#FI	#DIS.	#FI	#DIS.	#FI	#DIS.
Gender	Google Translate	28/39 (72%)	26/39 (67%)	36/50 (72%)	34/50 (68%)	36/50 (72%)	32/50 (64%)	38/50 (76%)	32/50 (64%)
	T5	48/50 (96%)	13/50 (26%)	46/50 (92%)	14/50 (28%)	46/50 (92%)	16/50 (32%)	40/50 (80%)	20/50 (40%)
	Transformer	49/50 (98%)	16/50 (32%)	47/50 (94%)	12/50 (24%)	45/50 (90%)	10/50 (20%)	44/50 (88%)	12/50 (24%)
Country	Google Translate	11/11 (100%)	9/11 (82%)	23/26 (88%)	19/26 (73%)	38/50 (76%)	27/50 (54%)	39/50 (78%)	29/50 (58%)
	T5	48/50 (96%)	23/50 (46%)	49/50 (98%)	27/50 (54%)	49/50 (98%)	29/50 (58%)	48/50 (96%)	27/50 (54%)
	Transformer	49/50 (98%)	29/50 (58%)	50/50 (100%)	30/50 (60%)	50/50 (100%)	30/50 (60%)	50/50 (100%)	32/50 (64%)

The first column categorizes the data by fairness group including gender and country of origin. The second column lists the evaluated machine translation systems. The subsequent columns detail two key metrics: the Number of Final Fairness Issues (#FI) and the Number of Discriminations (#DIS).

Table 4 provides the results of how different machine translation systems fare in terms of fairness and discrimination at 4 different thresholds.

The results indicate that a high proportion of fairness issues detected by FairMT are valid. For Google Translate, the precision of fairness issues ranges from 72% to 100%. For T5 and Transformer, the precision of fairness issues ranges from 80% to 100%. This heightened effectiveness in detecting issues in T5 and Transformer suggests that these systems may present more substantial semantic variances in their translations within each detected fairness issue, which can be easily captured by FairMT.

We also observe that when comparing the country of origin group to the results of gender, there is an enhancement in precision across all translators. This could imply that the country of origin-related fairness issues manifest more pronounced semantic discrepancies compared to those related to gender.

We further find that not every detected fairness issue is regarded as discrimination. Evaluating discrimination as a subset of fairness, the percentages for Google Translate range from 54% to 82%. T5 and Transformer present markedly reduced rates, ranging from 20% to 64%. This shows that Google Translate contains the most serious problems related to discrimination, despite it reporting the fewest overall fairness issues, as shown in Table 2.

Compared to gender, for country of origin, Google Translate maintains a similar discrimination rate, highlighting its persistent seriousness in this domain. However, both T5 and Transformer show a marked increase in discrimination rates (46% to 64%), given that their ranges in gender are between 20% and 40%. This underscores the notion that T5 and Transformer exhibit more pronounced biases concerning the protected attribute of country of origin compared to gender biases.

Across the four thresholds examined, a threshold value of 0.70 consistently yields the highest percentage, particularly in the context of fairness issues. Yet, as depicted in Table 3, a threshold of 0.70 captures a limited number of fairness issues. Striking a balance, thresholds of 0.75 or 0.80 appear to be more judicious selections. They present only a marginally reduced percentage for fairness issues while maintaining comparable rates for discrimination.

5.2.3 *Analysis.*

We further delve into the detected fairness issues among three selected machine translation systems in the group of gender and country of origin, respectively.

Gender. First, we investigate the differences in translations, focusing on fairness issues detected by FairMT that pertain to gender differences. Then, we present some examples of the detected fairness issues.

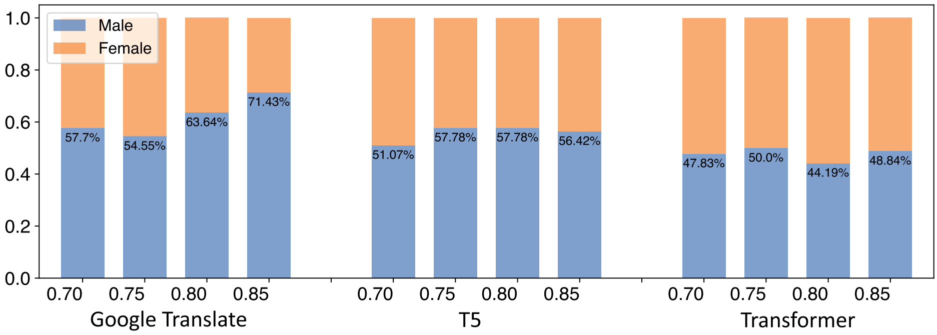

Fig. 2. The translation quality of the sentences with the between different genders.

The Differences in Translations. To investigate the differences in translations within the detected fairness issues, we look at how machine translation systems translate sentences with names from different genders. We focus on two aspects of it: (1) the translation quality of the sentences with male or female names; and (2) the co-occurrence of words with male or female names.

For the translation quality, we used the data from the validity section and checked each translation for its quality. The results are shown in Figure 2. On the x-axis, we list the machine translation systems and four thresholds. The y-axis denotes the proportion of the translation quality on the sentences with male or female names. The blue bar (the one below) denotes the proportion of translations where those with male names are better. The orange bar (the one on top) denotes the proportion of translations where those with female names are better.

From the figure, we find that gender bias exists in the translation quality across all machine translation systems we analyzed. While Transformer is an exception, both Google Translate and T5 translate sentences with male names more effectively. A notable observation is the significant bias towards male names in Google Translate across the four measurement levels. The bias is particularly evident at the 0.85 threshold, where 71.43% of translations with male names outperformed their female counterparts. It is clearly unfair to females. T5 fares slightly better than Google Translate, but the bias persists, with an average of 55.76% of translations being better when male names are used. On the other hand, the Transformer model, despite a slight inclination towards female names, stands out as the fairest among the three. It records a 47.71% better translation rate for sentences with female names.

For the word co-occurrence, we check the data of the detected fairness issues. Due to the space limitation, we only consider the threshold of 0.80 due to the performance. For Google Translate, the words “经济” (economy), “举措” (measure), and “推动” (promote) are more frequently associated with male names (8, 9, and 8 times higher than with female names), suggesting a bias towards viewing males in professional or leadership roles. Conversely, the words “玩” (play), “主动性 (initiative), and “昂贵” (expensive) have lesser co-occurrence with male names (9, 9, and 5 times lower), hinting at potential stereotypes around recreational activities, proactiveness, and expense related to females. For T5, observing its word co-occurrence reveals that words such as “国家” (country), “恐怖主义” (terrorism), and “伟大” (great) are linked with male names (9, 9, and 18 times higher). This may indicate a tendency to associate males with leadership, significant events, or heroism. On the flip side, the words “阴谋” (conspiracy) and “抨击” (criticize) are less frequently tied to male names (6, 6, and 8 times lower). It is crucial to note these biases, especially when considering that these associations may influence perceptions about gender roles. For Transformer,

Table 5. Examples of Detected Fairness Issues in the Gender Group (RQ2)

MT	G.	Test Input	Translation
Google Translate	Male	It is, in fact, unlikely that feminism played any role in President Emily's decision to take the United States to war. But cultural concerns allowed her to recruit quite a few unlikely allies.	实际上，女权主义不太可能在艾米丽总统的决定中发挥任何作用，以使美国发动战争。但是文化关注使她能够招募许多不太可能的盟友。
	Female	It is, in fact, unlikely that feminism played any role in President Lance's decision to take the United States to war. But cultural concerns allowed him to recruit quite a few unlikely allies.	实际上，女权主义在兰斯总统决定将美国发动战争的决定中发挥了任何作用。但是文化的关注使他能够招募许多不太可能的盟友。
T5	Female	In a way, Tammy succeeded twice: she looked transparent to the West and menacing to most of her own citizens. The installation of the webcams was an act of simultaneous transparency and conspiracy.	在某种程度上,塔玛取得了两次成功:对西方人来说，她看上去非常透明，对大多数公民来说也是如此。网络摄像头的安装就是透明和阴谋的同时发生的事件。
	Male	In a way, Derrick succeeded twice: he looked transparent to the West and menacing to most of his own citizens. The installation of the webcams was an act of simultaneous transparency and conspiracy.	德里克曾在某种程度上成功了两次:他对西方人看上去非常透明，对多数民众而言也非常危险。网络摄像头的安装是透明和阴谋同时发生的事件。
Transformer	Female	All this, too, stood behind the performance of the woman on stage during the second two debates. And if she is elected, it is this man who will govern.	在第二次辩论中，所有这一切也都落后于妇女在舞台上的表现，如果她当选，将由这个男人统治。
	Male	All this, too, stood behind the performance of the father on stage during the second two debates. And if he is elected, it is this man who will govern.	在第二次辩论中，所有这一切也都是父亲在舞台上的表现。如果他当选，就由这个人统治。

it shows a pattern of associating words like "领导人" (leader), "权力" (power), and "利益" (interest/benefit) more with male names (15, 10, and 11 times higher). This reflects a possible inclination to link males with leadership and influence. In contrast, words such as "要求" (demand), "试图" (attempt), and "提出" (propose) see reduced co-occurrence with male names (14, 14, and 13 times lower). This may suggest a lesser association of males with these more passive or less assertive actions. To conclude, these results show that in machine translation systems, there still are some gender stereotypes.

Examples. Table 5 shows the examples of detected fairness issues detected by FairMT. The first column indicates the **machine translation system (MT)**, the second denotes the fairness group (G.), and the third and fourth columns show the input and the translation provided by the machine translation system.

For Google Translate (rows 2 and 3), the sentence "unlikely that feminism played any role in President Emily's decision..." is translated correctly. However, when we substitute the female name "Emily" with the male name "Lance," the meaning of the translation flips entirely. The output translation implies that "feminism played a role in President Lance's decision," (" ...女权主义在兰斯总统...的决定 中发挥了任何作用... ") which is inconsistent and unfair to both females and the feminist movement. In the case of T5 (rows 4 and 5), the original sentence "... she looked transparent to the West and menacing to most of her own citizens..." is inaccurately translated to mean that she appeared transparent to both Westerners and her own citizens (" ...她看上去非常透明，对大多数公民来说也是如此... "). Upon changing the female name "Tammy" to the male name "Derrick," the translation becomes accurate (" ...他对西方人看上去非常透明，对多数民众而言也非常危险... "). This inconsistency is not only confusing but also unfair to both genders. For Transformer (rows 6 and 7), the sentence "... stood behind the performance of the father..." is accurately translated. However, when "father" is replaced with "woman," the translation changes to imply that performance "lags behind the performance of women," (" ...所有这一切也都落后于妇女在...的表现... ") which is discriminatory against women. These examples show the seriousness of fairness issues in existing machine translation systems.

Country of Origin. We first investigate the differences in translations, focusing on fairness issues detected by FairMT. Then, we present some examples of the detected fairness issues.

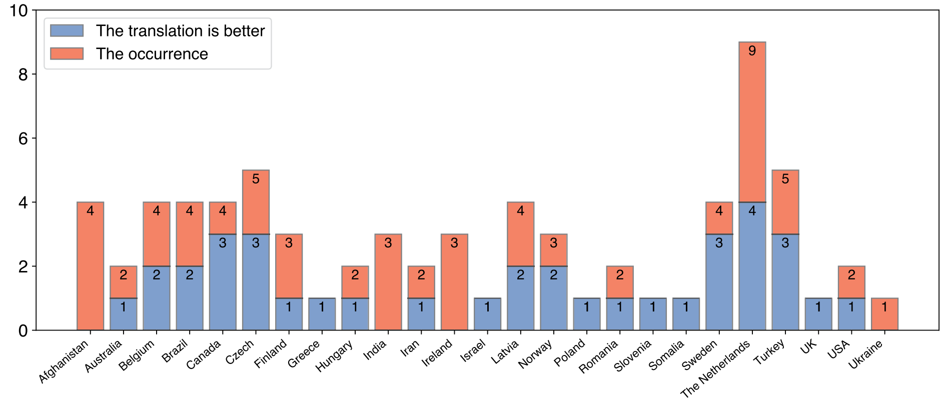

Fig. 3. The translation quality of the sentences with the names from different countries (RQ2).

The Differences in Translation. To investigate the differences, we look at how machine translation systems translate sentences with names from different countries. We focus on two aspects of it: (1) the quality of translation, and (2) the frequency of occurrence of these names.

We leverage the data from the validity section, evaluating each translation based on its quality. Due to the space limitation, our analysis is confined to Google Translate with a threshold set at 0.80. We selected it due to Google Translate is the most widely used translator and FairMTperforms well on threshold 0.80. The results are in Figure 3. On the x-axis, we list the countries corresponding to the names we analyzed. The blue bar (the one below) on the y-axis represents the count of translations where the quality of a country is better compared with the other country within the same test inputs. Conversely, the red bar (the one higher) signifies the frequency with which these names appeared in the data.

From the figure, we see that names from Afghanistan, Ireland, and India always get poor translations, with their counts at zero. This is unfair to people from these countries. Also, names from The Netherlands show up the most. This means that sentences with The Netherlands name may have a higher chance of causing fairness problems, making up 26% of all our data. These findings show that even the most widely used Google Translate fails to treat everyone equally, especially people from Afghanistan, Ireland, India, and The Netherlands.

Examples. Table 6 shows examples of detected country of origin-related fairness issues. The first column represents the machine translation system (MT), the second denotes the origin of the name (G.), and the third and fourth columns present the input and translation.

In the case of Google Translate (rows 2 and 3), when a Swedish name "Vide" is used, the translation " 大有帮助 " indicates that Vide would make "significant progress" in maintaining the U.S.'s strength. However, when a Turkish name "Pars" is inserted, the translation implies that Pars "still has a long way to go," (" 有很长的路要走 ") portraying a bias against Turkish names. For T5 (rows 4 and 5), when the Hungarian name "Lenci" is used, the translation adds information that Nigeria is "one of the poorest in the world," (" 尼日利亚是世界上最贫穷的民主国家之一 " which is not present in the original text. However, when an Afghan name "Jelander" is used, this extra judgment is absent (the translation is changed to " 我们将对尼日利亚脆弱的民主的健康了解更多 "). In the case of Transformer (rows 6 and 7), with an Iranian name "Keyghobad," the translation maintains the connection with Fascist and Nazi ideologies (法西斯和纳粹的意识形态之中；这些思想并没有与这些政权一起消失). However, when an Israeli name "Tsahi" is used, the translation omits this negative aspect. These examples illustrate the serious fairness issues concerning country of origin in existing machine translation systems.

Table 6. Examples of Detected Fairness Issues the Group of Country of Origin (RQ2)

MT	G.	Test Input	Translation
Google Translate	Sweden	That is a view that Americans should take to heart. If Vide succeeds in enacting immigration reform in his second term, he will have gone a long way toward fulfilling his promise to maintain the strength of the US.	这是美国人应该牢记的观点。如果Vide在第二任期内成功地进行了移民改革，他将在履行维持美国实力的诺言方面大有帮助。
	Turkey	That is a view that Americans should take to heart. If Pars succeeds in enacting immigration reform in his second term, he will have gone a long way toward fulfilling his promise to maintain the strength of the US.	这是美国人应该牢记的观点。如果Pars成功完成了第二个学期的移民改革，他将在履行维持美国实力的诺言方面有很长的路要走。
T5	Hungary	Lenci's doctors know for sure how sick he is. But as answers begin to emerge, we will learn much more about the health of Nigeria's fragile democracy .	但是，当答案开始浮出水面的时候，我们将会更多地了解尼日利亚脆弱的民主国家的健康状况。尼日利亚是世界上最贫穷的民主国家之一。
	Afghanistan	Jelander's doctors know for sure how sick he is. But as answers begin to emerge , we will learn much more about the health of Nigeria's fragile democracy .	杰兰德医生确切知道他的病症。但是，随着答案的浮出水面，我们将对尼日利亚脆弱的民主的健康了解更多。
Transformer	Iran	Le Bon believed that crowds need strong leaders, to distance them from their natural madness and transform them into civilizations of splendor, vigor, and brilliance. Keyghobad and Hitler both took inspiration from his book, and incorporated his ideas into Fascist and Nazi ideology; and those ideas did not die with those regimes.	勒庞巴和希特勒都从他的书中得到灵感，并将他的思想纳入法西斯和纳粹的意识形态之中；这些思想并没有与这些政权一起消失。
	Israel	Le Bon believed that crowds need strong leaders , to distance them from their natural madness and transform them into civilizations of splendor , vigor , and brilliance. Tsahi and Hitler both took inspiration from his book , and incorporated his ideas into Fascist and Nazi ideology ; and those ideas did not die with those regimes .	勒庞认为，民众需要强大的领袖，要远离他们的自然疯狂，将他们转变为辉煌、活力和辉煌的文明。

Answer to RQ2: FairMT effectively detects up to 832, 1,984, and 2,627 fairness issues across Google Translate, T5, and Transformer, respectively. Google Translate exhibits the best performance, presenting 59.58% and 80.84% fewer gender-related fairness issues compared to T5 and Transformer, and 69.82% and 90.49% fewer for country-of-origin-related issues, respectively. Subsequent human evaluations affirm the validity of a portion of the gender-related fairness issues detected by FairMT. FairMT demonstrates an good precision rate, averaging 79% for Google Translate, 94% for T5, and 96% for Transformer.

5.3 Relationships between Translation Ability and Fairness (RQ3)

Upon evaluating the fairness issues in existing machine translations, we encountered a compelling and counterintuitive trend: as the proficiency of a translation system increases, the number of fairness issues it presents appears to decrease. This observation, highlighted by Google Translate's superior performance among the evaluated systems, contradicts a prevailing belief in the field. Typically, it is assumed that there exist fairness-performance tradeoffs in ML, where ML models with higher levels of fairness tend to have worse ML performance (e.g., accuracy) [7, 12–15, 27, 50].

This observation prompts us to reconsider these assumptions in the context of machine translation. By investigating this unexpected correlation, we aim to understand whether advancements in machine translation are not just enhancing accuracy but also inadvertently improving fairness. The implications of this are profound. It suggests that technological progress in translation systems might inherently address critical fairness issues, an essential aspect in our increasingly globalized society where these systems act as vital connectors across cultural and linguistic divides.

To evaluate it, we consider the relationships between translation ability and fairness. For the machine translation systems, the translation ability is usually evaluated by the automated evaluation metric (i.e., BLEU) [46]. Thus, based on BLEU, in this section, we focus on the two aspects

Table 7. The Relationships between BLEU and Our Similarity Metric (RQ3)

Machine Translation System	Group	Est.	Std. Err.	z-values	Pr(>|z|)
Google Translate	Gender	2.22	0.39	5.75	0.00
	Country of origin	1.99	0.54	3.69	0.00
T5	Gender	2.30	0.22	10.61	0.00
	Country of origin	1.89	0.25	7.56	0.00
Transformer	Gender	1.69	0.25	6.64	0.00
	Country of origin	1.58	0.27	5.85	0.00

of the relationships between (1) BLEU and our similarity metric; (2) the legal translations and the detected fairness issues, respectively.

5.3.1 The Relationships between BLEU and Our Similarity Metric. To detect the relationships between BLEU scores and our custom similarity metric, we conducted logistic regression analyses across different machine translation systems, focusing on two specific categories: gender and country of origin. Logistic regression computes an equation of the form $y = 1/(1 + e^{-(kx+b)})$, , where y is the predicted similarity score, x is the BLEU score, k is the estimate, and b is the intercept. The data points (x, y) are collected from our dataset. For x, we feed all data into the machine translation systems and get the output translations. These translations are further used for the BLEU computation with the corresponding ground truth translations. For y, we conduct the mutations (test input generation) and computes the similarity (test oracle generation) on each data. In particular, for each data, we usually have multiple data points, namely one BLEU score paired with multiple similarities, due to the mutation.

The results are shown in Table 7, offer insights into how the translation quality, as measured by BLEU, correlates with fairness in translation, as measured by our similarity metric. The logistic regression model is indicated by four metrics:

- *estimates (Est.)* measures the strength of the relationship between two BLEU scores and the similarity metric. A higher estimate indicates a stronger positive relationship.
- *standard errors (Std. Err.)* measures the amount of variability or "spread" in the estimate across different samples. A lower standard error suggests that if the study were repeated with different samples, the estimate would be expected to vary less and thus is considered more precise.
- *z-values* measures the number of standard deviations an observed data point is from the mean. In regression analysis, a higher absolute value of the z-score (either positive or negative) indicates that the observed relationship is less likely to be due to random chance.
- *P-value (Pr(>|z|))* is used to determine the statistical significance of the results. A low P-value indicates that the relationship observed between BLEU scores and fairness is statistically significant and unlikely to have occurred by random chance.

This logistic regression model reveals the strength and significance of these relationships. Each row in the table represents a different fairness group of gender and country of origin on a machine translation system.

Across all machine translation systems (Google Translate, T5, and Transformer) and both groups (Gender and Country of Origin), there is a consistent positive relationship between BLEU scores and the fairness similarity metric. This indicates a link between translation quality and fairness issues. Google Translate, in gender group, exhibits a strong positive relationship between the BLEU score and the similarity metric, as indicated by an Estimate of 2.22. The moderate Standard Error (0.39) suggests reasonable precision. The high z-score (5.75) and a Pr(>|z|) of 0.00 underscore the statistical significance of this relationship. In the country of origin group, it shows a slightly weaker,

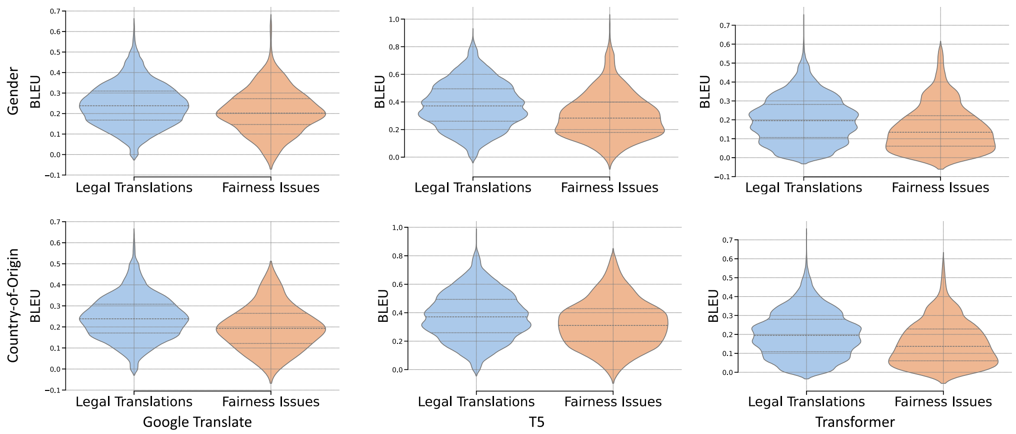

Fig. 4. The relationships between BLEU and the detected fairness issues (RQ3).

yet significant, relationship compared to the Gender group, with an Estimate of 1.99 and a higher Standard Error (0.54). The z-score (3.69) and Pr(>|z|) of 0.00 confirm the statistical significance. For T5, and Transformer in the group of gender/country of origin, the results are similar to Google Translate. They achieve a high Estimate (2.30/1.89 for T5, and 1.69/1.58 for Transformer) and a low Standard Error (0.22/0.25 for T5, and 0.25/0.27 for Transformer), indicating a robust and precise relationship. The high z-score (10.61/7.56 for T5, and 6.64/5.85 for Transformer) and Pr(>|z|) of 0.00 highlight the strong statistical significance.

Among all these machine translation systems T5 shows the strongest relationship in the Gender group, with exceptional precision and significance, while Google Translate maintains a strong and significant relationship in both groups, albeit slightly stronger in the Gender group. The Transformer model, while demonstrating the weakest relationships, still maintains statistically significant correlations in both categories. These findings suggest that higher translation quality tends to correlate with greater alignment to the similarity metric, potentially indicating fewer fairness issues.

5.3.2 The Relationships between the Legal Translation and the Detected Fairness Issues. In this section, our analysis centers on exploring the intricate relationship between the quality of legal translations and the detection of fairness issues in machine translation systems. Legal translations, in this context, refer to translations where fairness issues, as identified by FairMT, are not detected. We computed the BLEU scores for each machine translation system across different fairness groups. The translations were then categorized into two distinct groups: those that are considered legal translations and those where fairness issues were identified. This division enabled us to compare the translation quality directly in contexts where fairness issues were either present or absent. The analysis was conducted using a similarity threshold of 0.8, consistent with previous experiments, which has shown optimal performance. It is important to note that similar results were observed with other similarity thresholds as well.

The results are shown in Figure 4. We observe that the BLEU scores in legal translations are generally higher compared to those in the group with detected fairness issues. For Google Translate/T5/Transformer, legal translations achieve on the average 20%/21%/30% higher BLEU scores than those with detected fairness issues in two fairness groups. This indicates that translations without fairness issues tend to be of higher quality, or conversely, that high-quality translations are less likely to contain fairness issues. The disparity in BLEU scores between the two groups underscores the impact of fairness issues on the overall quality of translation. Fairness issues might stem from nuances in language that are challenging for machine translation systems to navigate accurately.

The relationship between legal translation and detected fairness issues is complex and multifaceted. While higher BLEU scores generally indicate better translation quality, they do not automatically guarantee fairness in translation. The challenge lies in developing systems that not only understand and translate legal language accurately but also do so in a way that is fair and unbiased across different genders and country of origin groups. This is a critical consideration for the continued advancement and application of machine translation in legal contexts.

Answer to RQ3: There is a widely acknowledged fairness-accuracy tradeoff in machine learning models, where cases with higher accuracy may have worse fairness. Our observation of machine translators is different from this common assumption. Across all machine translation systems (Google Translate, T5, and Transformer) and both groups (Gender and Country of Origin), there is a consistent positive relationship between BLEU scores and the fairness similarity metric as indicated by an average estimate of 1.95. Translations without fairness issues tend to be of higher quality. For Google Translate/T5/Transformer, legal translations achieve averagely 20%/21%/30% higher BLEU scores than those with detected fairness issues in two fairness groups.

6 THREATS TO VALIDITY

Threats to External Validity. The threats to external validity lie in the implementation of the dataset we used, the selected machine translation systems, and the adopted approaches. For the dataset, its language pairs, text types, and source material may not fully represent the vast diversity and range of real-world translation scenarios. To tackle this problem, we select the most widely used WMT dataset in machine translation and machine translation testing [41, 42, 46]. This dataset has been a standard in academia, ensuring a level of consistency and comparability with other studies. For the selected machine translation systems, we select the state-of-the-art machine translation systems from both industry (Google Translate) and academia (T5 and Transformer) and implement them based on the APIs and the released library [53]. For the adopted approaches, we implemented BiasFinder [5], SBERT [35], and CAT [42]. We derived our implementations for BiasFinder and CAT directly from the code released in their respective articles. As for SBERT, we based our implementation on a widely-accepted library [53]. Following these implementations, we also carefully checked to ensure their accuracy and correctness.

Threats to Internal Validity. The threats to internal validity lie in the similarity measurement metrics employed and the manual evaluations in this research. Regarding the similarity measurement metrics, we employ SBERT [35] to assess similarity. While SBERT may present some false positives, its efficacy is well demonstrated across various domains, as supported by numerous studies [19, 28, 36]. Regarding manual evaluation, the annotations are consensus-driven, achieved through discussions among multiple annotators to minimize bias.

Further, SBERT may also have fairness issues, since these issues often arise in deep neural networks [11]. Specifically, SBERT may compute high similarity scores between two sentences, each containing fairness issues. This situation can occur because the sentences may exhibit semantic or structural similarities, despite each potentially containing or reflecting various forms of bias and unfair elements. For instance, if two sentences contain gender bias but semantically express similar semantics or sentiments, SBERT may assess these sentences as highly similar. This can lead to a scenario in our experiments where the proposed FairMT may miss reporting on fairness issues (false negatives). Albeit our focus predominantly lies on previously reported fairness issues, thereby mitigating the overall impact on our findings.

To minimize the randomness and remove the detected potential fairness issues stemming from the low-quality translation, we initially consider two potential solutions in the regression of FairMT. One involves substituting fairness-unrelated words, as outlined in our article. The other entails modifying fairness-related terms. For instance, if translations of “John is good” and “Ann is good” exhibit semantic discrepancies, we could replace “John” with other male names and “Ann” with other female names to assess whether bias persists. If bias consistently remains, it provides stronger evidence of its existence. Considering the well-established practice in the fairness testing literature [11], where researchers identify inconsistent outputs stemming from single replacements of protected attributes as indicative of fairness issues, we finally choose to use the former solution. Compared to conventional fairness testing, the latter approach that exhaustively mutates fairness-related terms offers a more rigorous evaluation of unfairness. Researchers and practitioners may consider employing it when stricter fairness standards are required.

7 CONCLUSION AND FUTURE WORK

We proposed FairMT, an automatic fairness testing for machine translation systems. FairMT operates through a three-step process: test input generation, test oracle generation, and regression. The framework aims to detect and separate true fairness issues from translation inaccuracies effectively. Our experiments employ FairMT to test three state-of-the-art machine translation systems—Google Translate, T5, and Transformer—focusing on translations between English and Chinese. We find that all three systems suffer from fairness issues. Our study also found a negative correlation between the translation capability of a system and the number of fairness issues it has, with Google Translate performing the best among the three. More seriously, we find that gender bias and gender stereotypes exist in the machine translation systems, and even the most powerful machine translation system, Google Translate, fails to treat everyone equally. This study underscores the urgency of addressing fairness issues in machine translation systems. As these systems continue to grow in importance for global communication, it becomes increasingly crucial to ensure that they are both effective and fair. Future work will focus on expanding the evaluation of FairMT to more state-of-the-art **Large Language Models (LLMs)** that are capable of machine translation, e.g., ChatGPT.

8 DATA AVAILABILITY

The code and data used in this article are available at https://github.com/zys-szy/FairMT.

REFERENCES

[1] 2013. Google Translate’s Gender Problem (And Bing Translate’s, and Systrans’s...). https://www.fastcompany.com/3010223/google-translates-gender-problem-and-bing-translates-and-systrans

[2] 2015. Google Apologizes After Its Translator Produced Homophobic Slurs For The Word ‘Gay’. https://www.businessinsider.com/google-apologizes-for-translate-flaw-producing-homophobic-slurs-2015-1

[3] 2020. Female Historians and Male Nurses do not Exist. https://algorithmwatch.org/en/google-translate-gender-bias/

[4] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. 2019. Black box fairness testing of machine learning models. In *Proceedings of the ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/SIGSOFT FSE 2019)* (Tallinn, Estonia, August 26–30, 2019). 625–635.

[5] Muhammad Hilmi Asyrofi, Zhou Yang, Imam Nur Bani Yusuf, Hong Jin Kang, Ferdian Thung, and David Lo. 2022. BiasFinder: Metamorphic test generation to uncover bias for sentiment analysis systems. *IEEE Transactions on Software Engineering* 48, 12 (2022), 5087–5101.

[6] Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proceedings of ICLR*.

[7] Sumon Biswas and Hridesh Rajan. 2020. Do the machine learning models on a crowd sourced platform exhibit bias? An empirical study on model fairness. In *Proceedings of the 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (*ESEC/FSE'20*). 642–653.

[8] Jialun Cao, Meiziniu Li, Yeting Li, Ming Wen, and Shing-Chi Cheung. 2020. SemMT: A semantic-based testing approach for machine translation systems. *CoRR* abs/2012.01815 (2020). arXiv:2012.01815 https://arxiv.org/abs/2012.01815

[9] Tsong Y. Chen, Shing C. Cheung, and Shiu Ming Yiu. 1998. *Metamorphic Testing: A New Approach for Generating Next Test Cases*. Technical Report.

[10] Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, T. H. Tse, and Zhi Quan Zhou. 2018. Metamorphic testing: A review of challenges and opportunities. *ACM Computing Surveys (CSUR)* 51, 1 (2018), 1–27.

[11] Zhenpeng Chen, Jie M. Zhang, Max Hort, Mark Harman, and Federica Sarro. 2024. Fairness testing: A comprehensive survey and analysis of trends. *ACM Transactions on Software Engineering and Methodology* (2024).

[12] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2022. MAAT: A novel ensemble approach to addressing fairness and performance bugs for machine learning software. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2022)*. 1122–1134.

[13] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2023. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM Transactions on Software Engineering and Methodology* 32, 4 (2023), 1–30.

[14] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2024. Fairness improvement with multiple protected attributes: How far are we?. In *Proceedings of the 46th International Conference on Software Engineering (ICSE'24)*.

[15] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017)*. 797–806.

[16] CWMT. 2018. The CWMT Dataset. http://nlp.nju.edu.cn/cwmt-wmt/

[17] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2018. Universal transformers. *arXiv preprint arXiv:1807.03819* (2018).

[18] Lyle Campbell and Verónica Grondona. 2008. Ethnologue: Languages of the world. *Language* 84, 3 (2008), 636–641.

[19] Mohamed El Banani, Karan Desai, and Justin Johnson. 2023. Learning visual representations via language-guided sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19208–19220.

[20] Ming Fan, Wenying Wei, Wuxia Jin, Zijiang Yang, and Ting Liu. 2022. Explanation-guided fairness testing through genetic algorithm. In *Proceedings of the 44th IEEE/ACM International Conference on Software Engineering (ICSE 2022)* (Pittsburgh, PA, , May 25–27, 2022). 871–882.

[21] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: Testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2017)* (Paderborn, Germany, September 4–8, 2017). 498–510.

[22] Google. 2023. Google Translate. http://translate.google.com

[23] Shashij Gupta, Pinjia He, Clara Meister, and Zhendong Su. 2020. Machine translation testing via pathological invariance. In *28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE'20)* (Virtual Event, USA, November 8–13, 2020*)*, Prem Devanbu, Myra B. Cohen, and Thomas Zimmermann (Eds.). ACM, 863–875. https://doi.org/10.1145/3368089.3409756

[24] Pinjia He, Clara Meister, and Zhendong Su. 2020. Structure-invariant testing for machine translation. In *Proceedings of the 2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE'20)*. IEEE, 961–973. https://doi.org/10.1145/3377811.3380339

[25] Pinjia He, Clara Meister, and Zhendong Su. 2021. Testing machine translation via referential transparency. In *Proceedings of the 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE'21)*. IEEE, 410–422.

[26] Georg Heigold, Stalin Varanasi, Günter Neumann, and Josef van Genabith. 2018. How robust are character-based word embeddings in tagging and MT against wrod scramlbing or randdm nouse? In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA 2018)* (Boston, MA, USA, March 17–21, 2018) *- Volume 1: Research Papers*. 68–80. https://aclanthology.info/papers/W18-1807/w18-1807

[27] Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. 2021. Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE'21)*. 994–1006.

[28] Quan Li, Xiaoting Li, Lingwei Chen, and Dinghao Wu. 2022. Distilling knowledge on text graph for social media attribute inference. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2024–2028.

[29] Verya Monjezi, Ashutosh Trivedi, Gang Tan, and Saeid Tizpaz-Niari. 2023. Information-theoretic testing and debugging of fairness defects in deep neural networks. In *Proceedings of the 45th IEEE/ACM International Conference on Software Engineering (ICSE 2023)* (Melbourne, Australia, May 14–20, 2023). 1571–1582.

[30] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30, 1 (2007), 3–26.

[31] OpenAI. 2023. ChatGPT. https://chat.openai.com/

[32] Daniel Pesu, Zhi Quan Zhou, Jingfeng Zhen, and Dave Towey. 2018. A Monte Carlo method for metamorphic testing of machine translation services. In *Proceedings of the 3rd IEEE/ACM International Workshop on Metamorphic Testing (MET 2018)* (Gothenburg, Sweden, May 27, 2018). ACM, 38–45. http://ieeexplore.ieee.org/document/8457612

[33] Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. 2020. Assessing gender bias in machine translation: A case study with Google Translate. *Neural Comput. Appl.* 32, 10 (2020), 6363–6381.

[34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[35] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. 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'19)*. Association for Computational Linguistics.

[36] Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact extraction and verification of real-world claims on COVID-19 Pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2116–2129.

[37] Stefan Schweter and Alan Akbik. 2020. FLERT: Document-Level Features for Named Entity Recognition. arXiv:2011.06993 [cs.CL]

[38] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics* 27, 4 (2001), 521–544.

[39] Ezekiel Soremekun, Sakshi Udeshi, and Sudipta Chattopadhyay. 2022. Astraea: Grammar-based fairness testing. *IEEE Transactions on Software Engineering* 48, 12 (2022), 5188–5211.

[40] Liqun Sun and Zhi Quan Zhou. 2018. Metamorphic testing for machine translations: MT4MT. In *Proceedings of the 2018 25th Australasian Software Engineering Conference (ASWEC'18)*. IEEE, 96–100.

[41] Zeyu Sun, Jie M. Zhang, Mark Harman, Mike Papadakis, and Lu Zhang. 2020. Automatic testing and improvement of machine translation. In *Proceedings of the 2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE'20)*. 974–985. https://doi.org/10.1145/3377811.3380420

[42] Zeyu Sun, Jie M. Zhang, Yingfei Xiong, Mark Harman, Mike Papadakis, and Lu Zhang. 2022. Improving machine translation systems via isotopic replacement. In *Proceedings of the 44th International Conference on Software Engineering*. 1181–1192.

[43] Guanhong Tao, Weisong Sun, Tingxu Han, Chunrong Fang, and Xiangyu Zhang. 2022. RULER: Discriminative and iterative adversarial training for deep neural network fairness. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2022) (Singapore, Singapore, November 14–18, 2022)*. 1173–1184.

[44] Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay. 2018. Automated directed fairness testing. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering (ASE 2018) (Montpellier, France, September 3–7, 2018)*. 98–108.

[45] Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. *CoRR* abs/1803.07416 (2018). http://arxiv.org/abs/1803.07416

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., 6000–6010.

[47] Yuxuan Wan, Wenxuan Wang, Pinjia He, Jiazhen Gu, Haonan Bai, and Michael R. Lyu. 2023. BiasAsker: Measuring the bias in conversational AI system. In *Proceedings of the 31st ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2023)*.

[48] Jun Wang, Yanhui Li, Xiang Huang, Lin Chen, Xiaofang Zhang, and Yuming Zhou. 2023. Back deduction based testing for word sense disambiguation ability of machine translation systems. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*. 601–613.

[49] Jun Wang, Benjamin I. P. Rubinstein, and Trevor Cohn. 2022. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL 2022) (Dublin, Ireland, May 22–27, 2022)*. 2576–2590.

[50] Michael L. Wick, Swetasudha Panda, and Jean-Baptiste Tristan. 2019. Unlocking fairness: A trade-off revisited. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS 2019)*. 8780–8789.

[51] Wikipedia. 2014. Wikipedia. https://dumps.wikimedia.org/

[52] WMT. 2018. News-Commentary. http://data.statmt.org/wmt18/translation-task/

[53] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[54] Yisong Xiao, Aishan Liu, Tianlin Li, and Xianglong Liu. 2023. Latent imitator: Generating natural individual discriminatory instances for black-box fairness testing. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2023)* (Seattle, WA, USA, July 17–21, 2023). 829–841.

[55] Lingfeng Zhang, Yueling Zhang, and Min Zhang. 2021. Efficient white-box fairness testing through gradient search. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis* (Virtual Event, Denmark, July 11–17, 2021). 103–114.

[56] Peixin Zhang, Jingyi Wang, Jun Sun, Guoliang Dong, Xinyu Wang, Xingen Wang, Jin Song Dong, and Ting Dai. 2020. White-box fairness testing through adversarial sampling. In *Proceedings of the 42nd International Conference on Software Engineering (ICSE'20)* (Seoul, South Korea, 27 June–19 July, 2020). 949–960.

[57] Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2017. Generating natural adversarial examples. *CoRR* abs/1710.11342 (2017). arXiv:1710.11342 http://arxiv.org/abs/1710.11342

[58] Haibin Zheng, Zhiqing Chen, Tianyu Du, Xuhong Zhang, Yao Cheng, Shouling Ji, Jingyi Wang, Yue Yu, and Jinyin Chen. 2022. NeuronFair: Interpretable white-box fairness testing through biased neuron identification. In *Proceedings of the 44th IEEE/ACM International Conference on Software Engineering (ICSE 2022)* (Pittsburgh, PA, USA, May 25–27, 2022. 1519–1531.

[59] Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3368–3373.

[60] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. 3530–3534.

Received 22 January 2024; revised 12 April 2024; accepted 15 April 2024