Lost in Translationese? Reducing Translation Effect Using Abstract Meaning Representation

Shira Wein
Georgetown University
sw1158@georgetown.edu

Nathan Schneider
Georgetown University
nathan.schneider@georgetown.edu

Abstract

Translated texts bear several hallmarks distinct from texts originating in the language. Though individual translated texts are often fluent and preserve meaning, at a large scale, translated texts have statistical tendencies which distinguish them from text originally written in the language ("translationese") and can affect model performance. We frame the novel task of *translationese reduction* and hypothesize that Abstract Meaning Representation (AMR), a graph-based semantic representation which abstracts away from the surface form, can be used as an interlingua to reduce the amount of translationese in translated texts. By parsing English translations into an AMR and then generating text from that AMR, the result more closely resembles originally English text across three quantitative macro-level measures, without severely compromising fluency or adequacy. We compare our AMR-based approach against three other techniques based on machine translation or paraphrase generation. This work makes strides towards reducing translationese in text and highlights the utility of AMR as an interlingua.

1 Introduction

The term *translationese* (Gellerstam, 1986) describes the features unique to translated texts: the specific syntactic and semantic patterns found in human translations (Teich, 2003; Volansky et al., 2013). When the presence of translationese is not addressed in training or test sets, evaluation scores can be overinflated (Zhang and Toral, 2019; Graham et al., 2020; Wang et al., 2023a), model performance can be impacted (Yu et al., 2022; Ni et al., 2022), or system-generated output can be dispreferred by humans (Freitag et al., 2019). However, if used correctly, actively leveraging translated texts in language model training can lead to improved performance in machine translation systems (Parthasarathi et al., 2021; Kurokawa et al., 2009; Lembersky et al., 2012; Twitto et al., 2015).

Original translation: Now, however, he is to go before the courts once more because the public prosecutor is appealing.

Parsed AMR:

```
(c / contrast-01
    :ARG2 (g / go-02
        :ARG0 (h / he)
        :ARG4 (c2 / court)
        :mod (a / again
            :mod (o / once))
        :time (n / now)
        :ARG1-of (c3 / cause-01
            :ARG0 (a2 / appeal-02
                :ARG0 (p / person
                    :ARG0-of (p2 / prosecute-01)
                    :ARG1-of (p3 / public-02))))))
```

Generated sentence: But now he will go to court once again because the public prosecutor is appealing.

Figure 1: Example of our "parse-then-generate" approach to mitigating translationese, which involves first translating the sentence into an AMR and then generating back into a sentence.

Previous work has studied the characteristics and impact of translationese.[1] In this work, we set out to reduce the amount of translationese in human-translated text while preserving the meaning. This corresponds to a task of automatic *translationese reduction* for human translations (§3). This task is important given the effect of translationese in both training and test sets, and is relevant to automatic tools for post-editing translations.

We hypothesize that translationese can be reduced using a formal semantic representation as an interlingua, because the representation abstracts away from the surface form while ensuring the integrity and continuity of the core elements of meaning. Specifically, we explore the utility of the Abstract Meaning Representation (AMR; Banarescu et al., 2013) formalism as an interlingua/interme-

[1] Though the term "translationese" is still commonly used in NLP/MT, it is less commonly used in translation studies (Jimenez-Crespo, 2023). In this work, we use the term to refer to specific characteristics which may arise out of the translation process, not necessarily corresponding to unnaturalness in the text (Kunilovskaya and Lapshinova-Koltunski, 2019).

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 753–765

diate representation for this task. We introduce a "parse-then-generate" technique which takes a text affected by translationese, parses that text into an AMR, and then generates text which is more like original English from that AMR.

In addition to our proposed "parse-then-generate" technique leveraging AMR, we experiment with two additional promising techniques. First, given the similarity between our task of translationese reduction and paraphrase generation, we apply two paraphrase models (one T5-based and one BART-based) to translationese reduction. We suspect that these models should also reduce the effect of translation on the surface form and lead to reduced explicitation, which is a hallmark of translationese (Baker et al., 1993; Gellerstam, 1996). Next, given the promise of "back-translation" for this task and the distinct set of translationese features appearing in machine versus human translations (Bizzoni et al., 2020), we test whether back-translation using machine translation actually reduces the amount of human translationese (§4).

We assess the performance of each technique for translationese reduction through experimentation with three macro-level translationese metrics (§5), an automatic evaluation of meaning preservation using three NLG metrics (§6.1), a thorough human evaluation of both fluency and adequacy, and qualitative analysis of the output (§6.2).

While AMR generation does not produce perfectly fluent texts (as judged by human evaluators), we find that the AMR-based approach is the only method which aids in translationese reduction across all metrics while preserving sufficient adequacy and fluency, highlighting the promise of AMR as an interlingua. The code for the AMR parse-then-generate technique and our evaluation protocol is available at `https://github.com/shirawein/amr-translationese`.

2 Background on Translationese

Translated and non-translated text (originally written in that language) exhibit various differences referred to as "translationese" (Gellerstam, 1986). Translated text is often less lexically rich, has simpler constructions, exhibits explicitation, and demonstrates specific lexical and word order choices (Baker et al., 1993; Gellerstam, 1996). An example exhibiting translationese can be seen at the top of Figure 1. The presence of translationese is not necessarily indicative of a low-quality translation (Kunilovskaya and Lapshinova-Koltunski, 2019), and prior work has shown that human raters are not able to accurately predict whether text is translated or not (Tirkkonen-Condit, 2002; Wein, 2023).

Two basic types of translationese include: (1) interference from the source, such as the presence of syntactic patterns typical of the source language (Teich, 2003), and (2) over-normalizing to the target language, for example not translating abnormalities seen in the source text. The patterns and characteristics of translationese vary by mode and register, most notably if the translation is written or spoken (Bernardini et al., 2016); translationese found in human translations versus machine translations (MT) also exhibit different characteristics (Bizzoni et al., 2020).

Related work has also considered the impact and causes of translationese via investigating the algorithmic biases which lead to translationese in MT (Vanmassenhove et al., 2021), avoiding the influence of translationese in training and testing by means of translationese classifiers and zero-shot multilingual MT (Riley et al., 2020), and exploring the utility of word-by-word glosses in producing fluent translations (Pourdamghani et al., 2019).

Prior work has developed automatic classifiers of translationese, which detect whether the text exhibits translationese or not (Rabinovich and Wintner, 2015; Rabinovich et al., 2017; Pylypenko et al., 2021). A couple of studies have sought to counteract the effects of translationese. Contemporaneously to the present work, Jalota et al. (2023) evaluated translationese classifier accuracy before and after applying style transfer to translated texts. In a similar vein, Dutta Chowdhury et al. (2022) removed translationese implicitly encoded in vector embeddings (but did not produce a reduced-translationese version of the translated text). Our work is novel in that we (1) frame the task of translationese reduction as one which reduces the statistical patterns of translationese, while preserving meaning and fluency, (2) introduce three methods of translationese reduction, and (3) demonstrate on both quantitative and qualitative metrics that our AMR-based approach succeeds at reducing the presence of translationese.

3 Translationese Reduction

We undertake this task of automatic translationese reduction for English, where the input is a sentence

that has been translated into English and the output is a paraphrase that better resembles a sentence that originated in English. We do not assume access to the source sentence that was translated, or even to the source language.

We formulate the task of translationese reduction by proposing automatic metrics for diminishing the hallmarks of translationese, informed by prior work documenting the notable features of translated texts.

Importantly, fluency and adequacy must be preserved in the task of translationese reduction, as conveying the same meaning is paramount. Thus, the reduction of translationese hallmarks across various automatic metrics may not come at the cost of adequacy or fluency, and any viable method for translationese reduction needs to maintain these aspects of the text while reducing features of translationese.

In this work, we approach translationese reduction by first mapping the translated English into a meaning representation in order to abstract away from superficial aspects of expression that may be artifacts of the translation process. This meaning representation is intended as an intermediary, or "interlingua," between the two "dialects" of English: translationese and originally English text. For example, in Figure 1, we see that we start with a translation, parse the text into an AMR graph, and generate from that AMR graph a sentence more like original English text.

4 Methods

First, in §4.1, we introduce the data that we use for our experiments. Next, in the three subsections that follow, we outline the three approaches we develop to take on our task of translationese reduction: one using paraphrase generation models (§4.2); one using machine translation (§4.3); and the third approach using AMR as an interlingua (§4.4).[2]

4.1 Data

For our experiments, we use the *English corpus of Native, Non-native and Translated Texts* (ENNTT) (Nisioi et al., 2016), which is based on Europarl data (Koehn, 2005). ENNTT consists of three distinct (non-parallel) sets of data: translated text, text originally in English uttered by non-native speakers, and text originally in English uttered by native speakers. The translated texts are edited versions of the transcriptions, not real-time translations. To create the English Europarl proceedings/translated dataset, the spoken utterances were (1) transcribed, then (2) edited by the original speaker, then (3) translated by a human native speaker of English (Nisioi et al., 2016). Here, we use 2000 sentences from the translated and native datasets: 1000 translated sentences and 1000 originally English sentences uttered by native speakers. We use the originally English datasets to compare the part-of-speech values of translated English versus original English in §5.3.

4.2 Paraphrase Generation

Given that our goal of translationese reduction is a form of paraphrasing—producing a meaning-preserving alternative phrasing that better resembles originally English text—we experiment with two preexisting paraphrase models. We examine whether the produced paraphrases reduce the effects of translationese.

Para T5. The first is a T5-based paraphrase model (Vorobev and Kuznetsov, 2023b),[3] trained on the ChatGPT paraphrase dataset (Vorobev and Kuznetsov, 2023a). The model is based on the T5-base model and uses transfer learning to combine the benefits of the ChatGPT paraphrases and the paraphrases generated from this model. There are 420,000 items in the training data, with each consisting of a question and five paraphrases produced by ChatGPT.[4]

Para BART. The second paraphrase system uses BART (Lewis et al., 2019).[5] It was trained by fine-tuning a pretrained seq2seq (text2text) `bart-large` model on the Quora,[6] PAWS (Zhang et al., 2019), and MSR paraphrase corpora (Dolan and Brockett, 2005). The Quora Question Pair

[2] We also developed and experimented with an approach using syntactically controlled generation, adapting the model from Chen et al. (2019). However we found that this produced nonsensical output, as was the case for even the example generated sentences in Chen et al.'s (2019) paper. Thus, we have omitted this method from our results.

[3] https://huggingface.co/humarin/chatgpt_paraphraser_on_T5_base

[4] We use the AutoTokenizer pretrained from the `chatgpt_paraphraser_on_T5_base` model as well as the pretrained `chatgpt_paraphraser_on_T5_base` AutoModelForSeq2SeqLM.

[5] https://huggingface.co/eugenesiow/bart-paraphrase

[6] https://www.kaggle.com/c/quora-question-pairs

dataset consists of 404,290 rows; the PAWS (Paraphrase Adversaries from Word Scrambling) corpus consists of 751,450 rows; the MSR (Microsoft Research) paraphrase corpus consists of 5,800 pairs of sentences. All three datasets consist of paraphrase candidate pairs (the MSR dataset has a human annotation indicating whether the sentences are paraphrases).[7]

4.3 Round-Trip Machine Translation

The next approach uses round-trip machine translation through a second language. This approach is motivated by prior work which explored back- and forward-translation as a tool for identifying data which is original to the target language (not the source language). Riley et al. (2020) found that including back-translated data in translation models leads to a minor improvement in BLEU score. Round-trip machine translation has also been found to aid grammatical error correction under some conditions (Kementchedjhieva and Søgaard, 2023); this further motivates the use of machine translation in human translationese reduction, given that the features of translationese in human and machine translation are distinct (Bizzoni et al., 2020)). Because of prior work leveraging back- and forward-translation related to improving naturalness and identifying translationese, we suspect that this approach might aid in the reduction of characteristics of human translationese.

Using the EasyNMT package,[8] we take the original English text which is afflicted by translationese, translate it into French, and then translate the French back into English. We use French because it is a Europarl language and EN-FR machine translation is of high quality.

4.4 Abstract Meaning Representation

Our third and primary approach is to use semantic parsing to abstract away from the phrasing of the translation while maintaining meaning. We use the Abstract Meaning Representation formalism as the intermediate semantic representation because it captures the core elements of meaning while de-centering the specific phrasing associated with sentences. AMR encapsulates the meaning of a sentence in a rooted, directed graph. Each node in the graph corresponds to a semantic unit in the sentence, and is labeled with an entity or event type ("concept"). Edges between nodes reflect relationships between semantic units. We hypothesize that AMR is an especially good choice to serve as an interlingua in the reduction of translationese because it abstracts away from the surface form to isolate the semantic elements of the sentence. As function words, inflectional morphology, specific word order and word choice are not captured in the AMR, this could help deal with issues such as unnatural phrasing and promote lexical richness. Further, the abundance of work on text-to-AMR parsing and AMR-to-text generation means that the quality of output is relatively high compared to other semantic representations.

Upon the translated and (distinct) originally English sentences, we apply our "parse-then-generate" method: (1) parse the sentence into an AMR, then (2) from the parsed AMR, generate a sentence. This process is illustrated in Figure 1. We make use of the `amrlib`[9] BART-based text-to-graph AMR parser and T5-based graph-to-text generator.

To determine the effectiveness of using AMR as an interlingua to abstract away from translation effect, we apply three translationese metrics to see if the parsed-then-generated sentences have characteristics more similar to the originally English sentences than the translated sentences.

5 Measuring Translationese

Prior work has established several statistical properties of translated text (Volansky et al., 2013). Measures known to distinguish translations from non-translations include: (1) type-token ratio (TTR), (2) the presence of cohesive markers, and (3) unigram bag-of-part-of-speech (POS) tags. Note that while the metrics we apply here are informed by prior work both in natural language processing and translation studies, these metrics show a partial picture of the range of statistical patterns observed in translated texts. These are not "translation universals," per se, so much as they are statistical tendencies (Jimenez-Crespo, 2023) observed in prior work on features of translated texts.

We compare system outputs on these metrics, using the original translations as a baseline, to assess whether each system successfully mitigates the observed presence of translationese. In each subsection, we detail the metric as well as the results for each approach with that metric.

[7] We use the `BartForConditionalGeneration` pretrained model and the BARTTokenizer.

[8] https://github.com/UKPLab/EasyNMT

[9] https://github.com/bjascob/amrlib

	TTR (↑)	Cohesive Markers (↓)
Translations	0.0890	461
MT	0.0850	483
Para BART	0.1172 ✓	277 ✓
Para T5	0.0736	446 ✓
AMR P-then-G	0.1002 ✓	348 ✓

Table 1: Type-token Ratio (TTR) and number of cohesive markers for the 1000 translated sentences before and after using each of the translationese reduction methods. "MT" indicates MT back-translation and "AMR P-then-G" is an abbreviation for AMR Parse-then-Generate. ✓ indicates improvement over the baseline of the original translation.

5.1 Type-token Ratio

Type-token ratio (TTR), as used by Rabinovich et al. (2016), quantifies lexical richness. Lower TTR reflects *text simplification*, in which a sentence in the source language has fewer linguistically complex features upon translation into the target language (Blum and Levenston, 1978). Vanmassenhove et al. (2021) find a decrease in lexical richness in text affected by translationese.

Type-token ratio results can be found in Table 1. First, for our AMR parse-then-generate approach, the type-token ratio results point to success in reducing translationese. Type-token ratio, and thus lexical complexity, *increases* as expected once we apply our AMR parse-then-generate approach to the translated sentences. The AMR-based technique increases type-token ratio to 0.1002. The BART-based paraphrase model also successfully reduces the presence of translationese and leads to an even more drastic change, improving the type-token ratio to 0.1172.

However, we find that when applying the machine translation back-translation technique, type-token ratio decreases from 0.0890 to 0.0850, indicating further diminished linguistic complexity. Similarly, the T5-based paraphrase model diminishes lexical complexity and the type-token ratio is reduced to 0.0736.

5.2 Cohesive Markers

Cohesive markers are sentence transitions like "besides," "in other words," and "furthermore." They are often overused in translations (Rabinovich et al., 2016), consistent with the explicitation hypothesis (Blum-Kulka, 1986), which suggests that information implied or understood in an originally English text is often specified in translations. The presence of cohesive markers in the ENNTT corpus, which we use in this work, is investigated in Rabinovich et al. (2016). We would expect the presence of cohesive markers to decrease when successfully reducing the amount of translationese in a text.

In the case of cohesive markers, the MT back-translation technique again *exacerbates* translationese, with the number of cohesive markers increasing from 461 to 483. Both paraphrase models, on the other hand, reduce the number of cohesive markers: the T5-based paraphrase model produces a small decrease (from 461 to 446), while the BART-based paraphrase model leads to a much more drastic change (from 461 to 277).

The AMR parse-then-generate approach also successfully reduces the number of cohesive markers (from 461 to 348). Some cohesive markers are captured in the parsed AMRs (such as `contrast` being used to mark "however" in Figure 1), while cohesive markers which carry less meaning are not captured. This results in only information-carrying cohesive markers being included in the generated text, whereas less critical cohesive markers (which may be products of translationese) are omitted.

5.3 Unigram Bag-of-POS

Unigram bag-of-POS measures source interference on grammatical structure (Pylypenko et al., 2021; Volansky et al., 2013). As supported by the shining through hypothesis (Teich, 2003), the grammatical structure (as approximated by part-of-speech (POS) *n*-grams) of translationese-affected text should be more similar to that of the source language than text originally written in the target language. In order to collect part-of-speech tags for our test data, use the spaCy `en_core_web_sm` part-of-speech tagger.[10]

When using the AMR parse-then-generate approach, the unigram bag-of-POS results suggest that our approach decreases the proportion of key tags. Pylypenko et al. (2021) show that the POS tag relative frequency of ADV (adverbs) can be a predictor of the presence of translationese, perhaps as well as the relative frequency of determiners and adpositions. For all tags, the highest *number* of tags for most part-of-speech tags (12 out of 17) appears in the translated text. This is because the sentences are longer for the translated sentences than any other data, likely due to explicitation. The number

[10] https://spacy.io/models/en

	ADP	ADV	DET
Translations	0.1129	0.0433	0.0982
Originally English	0.1108	0.0389	0.0984
MT	0.1144	0.0413	0.1004
Para BART	0.1009	0.0457	0.0960
Para T5	0.1060	0.0333	0.0958
AMR P-then-G	0.1103	0.0419	0.0963

Table 2: Relative frequencies of three part-of-speech tags for the original translations and the generated text after application of each of our translationese reduction techniques. The relative frequencies of originally English text are also provided as a baseline.

of tokens in 1000 translated sentences is 32,596 in total; the number of tokens in 1000 translated parse-then-generated sentences is 27,958; for the 1000 originally English sentences the total number of tokens is 28,436; after parsing-then-generating the total number of tokens in the 1000 originally English sentences is 25,499.

The *proportion* of each POS tag in the dataset can be seen in Table 2. For three noteworthy tags which can predict whether a text is translated (adpositions, adverbs, and determiners), we see that using AMR as an interlingua decreases the proportion of these tags in the data, which is desired for ADP and ADV (but not for DET).

Similarly, the T5-based paraphrase model leads to a decrease in all three tags. The BART-based paraphrase model decreases the proportion of adpositions and determiners, but increases the proportion of adverbs. However, the MT back-translation output shows an increase in adpositions and determiners, and a decrease in adverbs.

5.4 Discussion of Translationese Metric Results

Our translationese metrics reveal that back-translation, via machine translation to French and then to English, does not reduce the amount of translationese in the human-translated texts, but rather *exacerbates* it for all three metrics. The T5-based paraphrase model similarly exacerbates the the amount of translationese except for on one metric, which is the unigram bag-of-POS. As such, quantitatively, we see an indication that these two methods are not effective techniques for translationese reduction.

On the other hand, we find that all three metrics point to a decrease in translationese with our AMR parse-then-generate approach. The same is true for the BART-based paraphrase model, which effectively reduces the amount of translationese on our metrics and shows the greatest reduction via type-token ratio and count of cohesive markers. Both the AMR parse-then-generate approach and the BART-based paraphrase model produce text more like the originally English text per the part-of-speech relative frequencies.

At this point, our results indicate that the BART-based paraphrase model or the AMR parse-then-generate technique may be a successful way to reduce translationese. In the next section, we examine whether adequacy and fluency are maintained or sacrificed using these methods of translationese reduction.

6 Evaluation of Fluency and Adequacy

Having established that macro-level indications of translationese are lessened by using either the BART-based paraphrase model or AMR as an interlingua, we now examine the quality of the generated sentences through the lenses of fluency and adequacy/meaning preservation. We report automatic metrics as well as results of a human evaluation study.

6.1 Automatic Metrics for Meaning Preservation

We use three metrics to automatically calculate meaning preservation via semantic similarity to the reference: BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020), and BERTscore (Zhang et al., 2020). The BERTscore version that we use relies on `roberta-large`. For COMET, we use the default `unbabel-comet` model.

As seen in Table 3, across all three metrics, MT back-translation has the highest semantic similarity score. This technique still fails to reduce translationese in the text (per §5).

The AMR parse-then-generate scores come in second highest for all three metrics. The improved

	BLEURT	COMET	BERTscore
MT	80.31 (1)	87.72 (1)	96.00 (1)
Para BART	60.05 (4)	74.31 (4)	94.02 (3)
Para T5	70.67 (3)	81.60 (3)	92.79 (4)
AMR P-then-G	75.81 (2)	84.95 (2)	94.89 (2)

Table 3: Average BLEURT, COMET, and BERTscore percentages and rankings (in parentheses) for the 1000 generated sentences from each of our three techniques for translationese reduction, compared against the original sentences as references.

Difference	Original (Translationese) Sentence	Sentence after parse-then-generate
Conciseness	"Mr President, I would firstly like to congratulate the rapporteur, Mr Koch, on his magnificent work and his positive cooperation with the Commission with regard to improving the texts and presenting this report and this proposal."	"First, I would like to congratulate Mr Koch for his magnificent work and his positive cooperation with the Commission in improving the texts and presenting this report and this proposal."
Cohesive Markers	"Most people, **however**, would like to live in the area in which they were born and raised, if they were given the chance to, **in other words**, if there was work there ."	"**But** if given the chance to do that (work there), most would like to live in the area where they were born and raised."
Word Order	"We note, first of all, that the committee considers the data, as presented in the Commission's annual report, to be in too aggregated a form to enable an in-depth evaluation of state aid policy which is simultaneously legitimate, sensitive to national interests and extensive in terms of compliance with the rules of competition, pursuant to the actual terms of the Treaty."	"First of all, we note that the committee considers the data presented in the Commission's annual report too aggregated to enable an in-depth evaluation of a legitimate state aid policy that is sensitive to national interests and is extensive in terms of compliance with competition rules within the actual terms of the Treaty."

Table 4: Examples of each of the three main differences we note in sentences before and after applying our AMR parse-then-generate method. The cohesive markers are bolded in the respective row.

Original (Translationese) Sentence	After BART-based paraphrase model
"Although there are now two Finnish channels and one Portuguese one, there is still no Dutch channel, which is what I had requested because Dutch people here like to be able to follow the news too when we are sent to this place of exile every month."	"Although there are now two Finnish channels and one Portuguese one, there is still no Dutch channel."
"Madam President, the presentation of the Prodi Commission's political programme for the whole legislature was initially a proposal by the Group of the Party of European Socialists which was unanimously approved by the Conference of Presidents in September and which was also explicitly accepted by President Prodi, who reiterated his commitment in his inaugural speech."	"Madam President, the presentation of the Prodi Commission's political programme for the whole legislature."

Table 5: Examples of brevity enforced by the BART-based paraphrase model, with the first example showing acceptable omission, and the second example demonstrating undue omission (with the sentence being incomplete).

naturalness of the AMR parse-then-generate output is also evident when examining input/output pairs. Three major differences we observed after applying the AMR parse-then-generate techniques include (1) change in word order, (2) reduction in cohesive markers, and (3) added conciseness. An example of each of these three differences can be seen in Table 4.

Both paraphrase models show substantially decreased semantic similarity, suggesting they may not accurately convey the meaning of the original sentence. Even the BART-based paraphrase model, which effectively reduced translationese across all three translationese metrics, suffers from low automatic metric scores, reaching a BLEURT score of 60.05 and a COMET score of 74.31. The BART-based paraphrase model has a BERTscore higher than the T5-based paraphrase model, though all four of the BERTscores are quite high and close to each other. The low scores are likely due to the the paraphrase models emphasizing brevity so much that key information is being discarded. For example, the first item in Table 5 shows an acceptable form of brevity, where the omitted content is not essential to reflecting the meaning of the original sentence, whereas the second example unduly cuts out relevant content and is not a complete sentence. The average sentence length for the BART-based paraphrase model is 15.07 tokens, whereas the average sentence length for the original (translationese) sentences is 31.33 tokens.[11]

Thus, the AMR-based technique strikes the best balance between translationese reduction and meaning preservation when assessed via automatic metrics.

6.2 Human Evaluation

Finally, we assess adequacy and fluency of the system output through a human evaluation study. We collect two judgments per item on 75 sets of items, where each set of items consists of all system outputs associated with one original translationese sentence. For adequacy, there were five sentences per item, and for fluency there were six sentences per item, because the original text was also judged. In total, this amounts to 1,650 total judgments (75 ×

[11] The average sentence length for the AMR parse-then-generate approach is 24.52 tokens; average sentence length for the T5-based paraphrase model is 22.42; average sentence length for the MT back-translation approach is 27.35.

	Avg Adequacy	Avg Fluency
MT	3.59 (1)	3.35 (2)
Para BART	2.45 (4)	1.91 (5)
Para T5	2.97 (3)	3.39 (1)
AMR P-then-G	3.34 (2)	2.76 (4)
Originals	N/A	3.19 (3)

Table 6: Average adequacy and fluency scores (and their rankings in parentheses) from our human evaluations on 75 sentence sets, comprising 1,650 total judgments. Originals were used as references in adequacy judgments.

5 = 375 adequacy judgments, doubly annotated = 750, plus 75 × 6 = 450 fluency judgments doubly annotated, equals 900).

12 annotators participated in total and each annotator judged 25 sets. Adequacy and fluency judgments were collected separately and by different annotators. All annotators were either Computer Science or Linguistics graduate students, and all annotators of fluency were native speakers of English. The order of the system output was randomized, such that no individual system would always appear first in the survey.

The annotators were asked to judge fluency on a scale from 1–4 and adequacy on a scale from 1–4 in reference to the original translationese-afflicted sentence. For fluency, a score of 1 corresponds with text which is "nonsensical", a score 2 is assigned for text which is "poor" and has many errors which make the text hard to understand, a score of 3 indicates that the quality of the text is "good" and largely understandable with few errors, and a score of 4 is for "flawless" text—perfectly formed English with no mistakes. For adequacy, text which has "no meaning preservation" and is completely unrelated to the reference receives a score of 1, text which exhibits "some meaning preservation" corresponds with a score of 2, text which has "most" of the same meaning as the reference gets a score of 3, and a score which conveys "all" of the same meaning receives a scores of 4.

The results of this study can be seen in Table 6. Inter-annotator agreement via Spearman's correlation is 0.5 for both fluency and adequacy, suggesting moderate agreement, and the automatic metrics of fluency and adequacy show the same pattern as the human evaluation.

Generally, we find that the MT back-translation and AMR parse-then-generate approaches achieve the highest adequacy, as indicated by the automatic metrics (Table 3). While the T5-based paraphrase model output is highly fluent, its adequacy is low, and does not effectively reduce translationese per our prior translationese metrics (§5). The AMR parse-then-generate output suffers from a lower degree of fluency than the MT back-translation and T5-based paraphrase approaches, though the AMR-based output is still sufficiently fluent (as judged qualitatively and via automatic metrics) to ensure readability and meaningfulness. Further progress on AMR-to-text generation models will enable more fluent output.

Additionally, it is worth noting that fluency is low in the human evaluation even for the human-produced originals. As indicated in annotators' comments, this low fluency is likely due to the domain being European Parliament proceedings, which can be complicated for lay people to comprehend (even as our fluency annotators were all native speakers of English).

6.3 Tradeoff between Translationese Reduction and Maintaining Fluency/Adequacy

Our results reveal the tradeoff between reducing the presence of translationese, while maintaining fluency and adequacy. Given that the goal is translationese reduction in text, our AMR-based approach is best suited for this task. Across three metrics, we demonstrate the utility of AMR in making translated texts more similar to originally English texts. The AMR parse-then-generate method doesn't perfectly maintain fluency, but based on the automatic metrics and human judgments, still achieves fluency only a bit below that of the original human utterances. Importantly, adequacy is maintained by the AMR parse-then-generate approach, indicating that information is not lost by using AMR as an interlingua, and suggesting that humans perhaps disprefer the phrasing of the AMR output, while it is still accurately conveying the necessary information.

7 Related Tasks

Our task of translationese reduction on human-translated text is related to the tasks of style transfer, grammatical error correction, paraphrase generation, text simplification, and automatic post-editing, because all of these aim to edit text after generation or produce new text with the same meaning as other text.

Style transfer and grammatical error correction

aim to control features of generated text. Style transfer can control, for example, whether the style is modern or classical, honorific or non-honorific, or conforms to European or Brazilian Portuguese (Wang et al., 2023b). Style transfer considers what type of style the generated/translated text takes on, not whether the text has broader features of translationese. Recent work on style transfer has leveraged AMR as an interlingua (Jangra et al., 2022). Grammatical error correction removes errors from text (Wang et al., 2021) and aims for fluency, but even error-free fluent text can exhibit features of translationese, such as the source language shining through (§6).

Paraphrase generation is the task of producing sentences which have essentially the same meaning but different syntax and/or word choice (Zhou and Bhat, 2021). Huang et al. (2022) use AMR to control the semantics of generated paraphrases. Similarly, paraphrase detection determines whether one sentence has the same meaning as another. Issa et al. (2018) combine AMR parses with latent semantic analysis to compare two sentences and identify whether they are paraphrases.

Text simplification aims to make text more readable and easier to process (Chandrasekar et al., 1996). Research on this task has employed a variety of neural models (Nisioi et al., 2017).

While in this work we focus on adjusting human translations, a related goal might be to reduce translationese in machine translation output. Reducing translationese in machine translations is distinct from automatic post-editing,[12] not only because modern automatic post-editing requires the use of both the source sentence and the translation (while we do not assume access to any information other than the translation we aim to alter) (Chollampatt et al., 2020), but more importantly because post-editing exhibits a heightened degree of translationese (Toral, 2019).

Other research at the intersection of AMR and translation has used AMR to improve neural machine translation, unrelated to translationese (Song et al., 2019; Nguyen et al., 2021; Li and Flanigan, 2022), and framed AMR generation as a machine translation problem (Pust et al., 2015; Castro Ferreira et al., 2017).

[12]Human post-editing involves humans looking at generated translations and altering them for increased fluency/quality; automatic post-editing aims to automate this process (do Carmo et al., 2021).

8 Conclusion

In this work, we investigated the task of translationese reduction and introduced three methods for this task. Our automatic metrics of translationese indicate that the task of translationese reduction is complicated, because we want translationese to be reduced without sacrificing fluency or adequacy (this tradeoff is discussed in §6.3). Overall, we find that using AMR as an interlingua aids in translationese reduction. By contrast, a BART-based paraphrase model is even more effective at reducing translationese, but dramatically over-summarizes, severely harming adequacy and fluency. The T5-based paraphrase model and MT back-translation approach do not show promise for this task.

Our findings suggest that translationese reduction could be performed as an additional step after translating to make the text more like originally English text, and provides further indication that AMR can serve as an interlingua for a range of tasks which require abstracting away from specific language features (cf. Xue et al., 2014; Wein et al., 2022; Song et al., 2019; Li and Flanigan, 2022).

Limitations

Despite much work on text-to-AMR parsing and AMR-to-text generation, there is of course some amount of error introduced in our parse-then-generate method. We find in our results that the meaning is preserved, and while fluency is a bit lower, additional progress on AMR-to-text generation research will likely enable further fluency in the end result of using AMR as an interlingua.

Future work may explore the applicability of our methods to languages other than English and additional domains. Further, because we have used European Parliament data in this experimentation, all of the source languages are European languages, and translationese has different features depending on the source language (Koppel and Ordan, 2011). AMR (and parsers/generators for it) has also been adapted to a number of languages other than English, so in principle it is possible to apply the same technique to different types of texts affected by translationese (Wein and Schneider, 2022). While we have not yet examined the downstream effect of applying our approach, this would be a promising avenue for future work.

Acknowledgements

This work is supported in part by a Clare Boothe Luce Scholarship and NSF award IIS-2144881. We thank Sireesh Gururaja and anonymous reviewers for their feedback. Thank you to the following people for supplying human judgments of fluency and adequacy: Chao-Chin Liu, Sajad Sotudeh, Jianan Su, Ke Lin, Xiulin Yang, Laasya Bangalore, Devika Tiwari, Autumn Toney-Wails, Rahel Fainchtein, Ryan Wails, Samuel King, and Thomas Lupicki.

References

Mona Baker, Gill Francis, and Elena Tognini-Bonelli. 1993. Corpus linguistics and translation studies: Implications and applications, chapter 2.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Silvia Bernardini, Adriano Ferraresi, and Maja Miličević. 2016. From EPIC to EPTIC—exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies*, 28(1):61–86.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Shoshana Blum and Eddie A Levenston. 1978. Universals of lexical simplification. *Language learning*, 28(2):399–415.

Shoshana Blum-Kulka. 1986. Shifts of cohesion and coherence in translation. in interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies, edited by Juliane House and Shoshana Blum-Kulka, 17–36. *Tübingen: Narr*.

Thiago Castro Ferreira, Iacer Calixto, Sander Wubben, and Emiel Krahmer. 2017. Linguistic realisation as machine translation: Comparing different MT models for AMR-to-text generation. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 1–10, Santiago de Compostela, Spain. Association for Computational Linguistics.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable paraphrase generation with a syntactic exemplar. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5972–5984, Florence, Italy. Association for Computational Linguistics.

Shamil Chollampatt, Raymond Hendy Susanto, Liling Tan, and Ewa Szymanska. 2020. Can automatic post-editing improve NMT? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2736–2746, Online. Association for Computational Linguistics.

Félix do Carmo, Dimitar Shterionov, Joss Moorkens, Joachim Wagner, Murhaf Hossari, Eric Paquin, Dag Schmidtke, Declan Groves, and Andy Way. 2021. A review of the state-of-the-art in automatic post-editing. *Machine Translation*, 35:101–143.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*. Asia Federation of Natural Language Processing.

Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef Genabith. 2022. Towards debiasing translation artifacts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3983–3991, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. *Translation studies in Scandinavia*, 1:88–95.

Martin Gellerstam. 1996. Translations as a source for cross-linguistic studies. *Lund studies in English*, 88:53–62.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Kuan-Hao Huang, Varun Iyer, Anoop Kumar, Sriram Venkatapathy, Kai-Wei Chang, and Aram Galstyan.

2022. Unsupervised syntactically controlled paraphrase generation with Abstract Meaning Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1547–1554, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fuad Issa, Marco Damonte, Shay B. Cohen, Xiaohui Yan, and Yi Chang. 2018. Abstract Meaning Representation for paraphrase detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 442–452, New Orleans, Louisiana. Association for Computational Linguistics.

Rricha Jalota, Koel Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2023. Translating away translationese without parallel data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7086–7100, Singapore. Association for Computational Linguistics.

Anubhav Jangra, Preksha Nema, and Aravindan Raghuveer. 2022. T-STAR: Truthful style transfer using AMR graph as intermediate representation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8805–8825, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Miguel A. Jimenez-Crespo. 2023. "Translationese" (and "post-editese"?) no more: on importing fuzzy conceptual tools from Translation Studies in MT research. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 261–268, Tampere, Finland. European Association for Machine Translation.

Yova Kementchedjhieva and Anders Søgaard. 2023. Grammatical error correction through round-trip machine translation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2163–2170, Dubrovnik, Croatia. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Maria Kunilovskaya and Ekaterina Lapshinova-Koltunski. 2019. Translationese features as indicators of quality in English-Russian human translation. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 47–56, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of Machine Translation Summit XII: Papers*, Ottawa, Canada.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265, Avignon, France. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Changmao Li and Jeffrey Flanigan. 2022. Improving neural machine translation with the Abstract Meaning Representation by combining graph and sequence transformers. In *Proceedings of the 2nd Workshop on Deep Learning on Graphs for Natural Language Processing (DLG4NLP 2022)*, pages 12–21, Seattle, Washington. Association for Computational Linguistics.

Long HB Nguyen, Viet H Pham, and Dien Dinh. 2021. Improving neural machine translation with AMR semantic graphs. *Mathematical Problems in Engineering*, 2021.

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Original or translated? a causal analysis of the impact of translationese on machine translation performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics.

Sergiu Nisioi, Ella Rabinovich, Liviu P. Dinu, and Shuly Wintner. 2016. A corpus of native, non-native and translated texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4197–4201, Portorož, Slovenia. European Language Resources Association (ELRA).

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Prasanna Parthasarathi, Koustuv Sinha, Joelle Pineau, and Adina Williams. 2021. Sometimes we want ungrammatical translations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3205–3227, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nima Pourdamghani, Nada Aldarrab, Marjan Ghazvininejad, Kevin Knight, and Jonathan May. 2019. Translating translationese: A two-step approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3057–3062, Florence, Italy. Association for Computational Linguistics.

Michael Pust, Ulf Hermjakob, Kevin Knight, Daniel Marcu, and Jonathan May. 2015. Parsing English into Abstract Meaning Representation using syntax-based machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1143–1154, Lisbon, Portugal. Association for Computational Linguistics.

Daria Pylypenko, Kwabena Amponsah-Kaakyire, Koel Dutta Chowdhury, Josef van Genabith, and Cristina España-Bonet. 2021. Comparing feature-engineering and feature-learning approaches for multilingual translationese classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8596–8611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. 2016. On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1881, Berlin, Germany. Association for Computational Linguistics.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

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

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using AMR. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Elke Teich. 2003. *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. De Gruyter Mouton, Berlin, Boston.

Sonja Tirkkonen-Condit. 2002. Translationese—a myth or an empirical fact?: A study into the linguistic identifiability of translated language. *Target. International Journal of Translation Studies*, 14(2):207–220.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Naama Twitto, Noam Ordan, and Shuly Wintner. 2015. Statistical machine translation with automatic identification of translationese. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 47–57, Lisbon, Portugal. Association for Computational Linguistics.

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

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Vladimir Vorobev and Maxim Kuznetsov. 2023a. ChatGPT paraphrases dataset.

Vladimir Vorobev and Maxim Kuznetsov. 2023b. A paraphrasing model based on ChatGPT paraphrases.

Jiaan Wang, Fandong Meng, Yunlong Liang, Tingyi Zhang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2023a. Understanding translationese in cross-lingual summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3837–3849, Singapore. Association for Computational Linguistics.

Yifan Wang, Zewei Sun, Shanbo Cheng, Weiguo Zheng, and Mingxuan Wang. 2023b. Controlling styles in neural machine translation with activation prompt. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2606–2620, Toronto, Canada. Association for Computational Linguistics.

Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A comprehensive survey of grammatical error correction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–51.

Shira Wein. 2023. Human raters cannot distinguish English translations from original English texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12266–12272, Singapore. Association for Computational Linguistics.

Shira Wein, Wai Ching Leung, Yifu Mu, and Nathan Schneider. 2022. Effect of source language on AMR structure. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 97–102, Marseille, France. European Language Resources Association.

Shira Wein and Nathan Schneider. 2022. Accounting for language effect in the evaluation of cross-lingual AMR parsers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3824–3834, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nianwen Xue, Ondřej Bojar, Jan Hajič, Martha Palmer, Zdeňka Urešová, and Xiuhong Zhang. 2014. Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1765–1772, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sicheng Yu, Qianru Sun, Hao Zhang, and Jing Jiang. 2022. Translate-train embracing translationese artifacts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 362–370, Dublin, Ireland. Association for Computational Linguistics.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proc. of ICLR*, Online.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.