Investigating the isometric behaviour of Neural Machine Translation models on binary semantic equivalence spaces

Atreya Shankar
Cognitive Systems, University of Potsdam
Department of Computational Linguistics, University of Zürich
atreya.shankar@{uni-potsdam.de,uzh.ch}

Abstract

Isometry is defined mathematically as a distance-preserving transformation between two metric spaces. In this research, we assume that well-performing Neural Machine Translation (NMT) models function approximately isometrically on semantic metric spaces and hypothesize that the frequency of such isometric behaviour correlates positively with general model performance. We conduct our investigation by using two NMT models of varying performance to translate semantically-equivalent German paraphrases, based off diverse WMT19 test data references, to English. We simplify the notion of semantic metric spaces into probabilistic binary semantic equivalence spaces and compute these using three transformer language models fine-tuned on Google's PAWS-X paraphrase detection task. By analyzing the paraphrase detection outputs, we show that the frequency of semantically isometric behaviour indeed correlates positively with general model performance. With our final results, we provide evidence both for and against claims made by other studies on automatic sequence evaluation metrics and NMT models' robustness to adversarial paraphrases.

1 Introduction

Isometry is defined mathematically as a distance-preserving transformation between two metric spaces (Coxeter, 1961). In this research, we view Neural Machine Translation (NMT) models from the perspective of semantic isometry and assume that well-performing NMT models function approximately isometrically on semantic metric spaces. That is to say, if two sentences are semantically equivalent on the source side, they should remain semantically equivalent after translation on the target side given a well-performing NMT model. A simplified illustration of isometry in higher dimensional functional spaces can be seen in Figure 1.

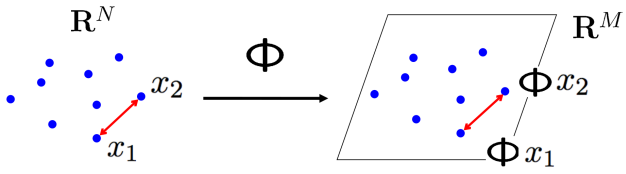

Figure 1: Illustration of isometry in higher dimensional functional transformations (Hegde et al., 2015)

We hypothesize that the frequency of such semantically isometric behaviour correlates positively with general model performance.

In order to conduct our investigation, we start by acquiring semantically equivalent German paraphrases of WMT19 legacy and additional test references from Freitag et al. (2020). Next, we utilize two NMT models of varying performance, specifically the SOTA FAIR's WMT19 winning single transformer (Ng et al., 2019) and the non-SOTA Scaling NMT WMT16 Transformer (Ott et al., 2018), to translate the aforementioned paraphrases to English.

Next, we simplify the notion of semantic metric spaces into probabilistic binary semantic equivalence spaces and utilize three well-performing paraphrase detection models to approximate these spaces in the NMT models' inputs and translations. These paraphrase detection models are based off the $\text{mBERT}_{\text{Base}}$ (Devlin et al., 2019), $\text{XLM-R}_{\text{Base}}$ (Conneau et al., 2019) and $\text{XLM-R}_{\text{Large}}$ (Conneau et al., 2019) pre-trained multilingual language models; which are correspondingly fine-tuned on Google's PAWS-X paraphrase detection task (Yang et al., 2019; Hu et al., 2020).

Using the outputs of the paraphrase detection models, we show that the frequency of semantically isometric behaviour correlates positively with general model performance. With our final results, we provide evidence both for and against claims made by other studies on automatic sequence evaluation

metrics and NMT models' robustness to adversarial paraphrases. We release our latest models and source code in our public GitHub repository[1].

2 Isometry and approximations

Isometry in the context of NMT and semantic metric spaces can be *exactly* expressed as follows; where $s_i \in \mathbb{R}^{V \times N}$ refers to an input sentence's tokenized matrix form for vocabulary size V and maximum sentence length N, $f : \mathbb{R}^{V \times N} \rightarrow \mathbb{R}^{V' \times N'}$ refers to the NMT model's inference function which translates sentences from language X to Y and $D_L : \mathbb{R}^{V \times 2N} \rightarrow \mathbb{R}_+$ refers to a semantic distance metric function for language L corresponding to the language of the respective sentences:

$$D_X(s_1, s_2) = D_Y(f(s_1), f(s_2)) \tag{1}$$

While elegant, this representation of isometry is problematic for two key reasons.

1. Exact isometry may not be a practical condition given real-life data instances with stochastic noise.
2. Constructing continuous semantic metric spaces from discrete textual data is a difficult task and is in itself a developing field of research (Cer et al., 2017; Michel et al., 2019).

2.1 Approximate isometry

To address the first issue, we loosen the constraints of exact isometry to *approximate* isometry:

$$D_X(s_1, s_2) \approx D_Y(f(s_1), f(s_2)) \tag{2}$$

With this approximation, we can simplify the isometric relationship further into a binary semantic equivalence function $S_L : \mathbb{R}^{V \times 2N} \rightarrow \{0, 1\}$, which compresses semantic distance metrics to semantic equivalence ($S_L = 1$) and inequivalence ($S_L = 0$) depending on some variable threshold $\delta_L \in \mathbb{R}_+$:

$$S_L(s_1, s_2) = \begin{cases} 1, & D_L(s_1, s_2) \leq \delta_L \\ 0, & D_L(s_1, s_2) > \delta_L \end{cases} \tag{3}$$

It is worth noting that the formulation in S_L is more meaningful for inferring isometry from semantic equivalence ($S_L = 1$) than from semantic inequivalence ($S_L = 0$), due to the presence of a tighter bound for the former than the latter.

[1] https://github.com/atreyasha/semantic-isometry-nmt

2.2 Probabilistic semantic equivalence spaces

To address the second issue, we effectively delegate away the actual computation of a semantic distance metric and convert this into a probabilistic process; with a new definition for S_L below given a probability threshold ϵ with a typical value of 0.5. This reformulation allows for the utility of statistical paraphrase detection models without explicit computation of semantic metric spaces:

$$S_L(s_1, s_2) = \begin{cases} 1, & P\big(D_L(s_1, s_2) \leq \delta_L\big) \geq \epsilon \\ 0, & P\big(D_L(s_1, s_2) \leq \delta_L\big) < \epsilon \end{cases} \tag{4}$$

With the aforementioned simplifications, we now re-write our equation for approximate isometry as follows:

$$S_X(s_1, s_2) = S_Y(f(s_1), f(s_2)) \tag{5}$$

For brevity, we use the terms *isometry* and *approximate isometry* interchangeably.

3 Related work

Based on a survey of recent literature in Natural Language Processing (NLP) and NMT, we were unable to find explicitly similar studies to our research. However, we would argue that the closest field in NLP to this research would be *adversarial paraphrasing*.

Michel et al. (2019) describes adversarial paraphrasing in the purview of machine translation as constructing paraphrases that are *"meaning preserving on the source-side, but meaning-destroying on the target-side"*. For the sake of comparison, we would mildly *paraphrase* this description of adversarial paraphrasing to *"the process of perturbing an input sentence such that it is semantically equivalent on the source-side, but semantically inequivalent on the target-side"*.

In this sense, the study of adversarial paraphrasing in machine translation could be interpreted as a probe into semantic *anisometry* of NMT models, compared to our research which would be a probe into semantic isometry of NMT models. Therefore adversarial paraphrasing, while having the opposite intent, is still highly similar to our research.

Michel et al. (2019): This research lays out the framework for evaluating adversarial perturbations in sequence-to-sequence models. Additionally, this

research compared three automatic sequence evaluation metrics, specifically BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and chrF$_2$ (Popović, 2015), against human judgment for evaluating semantic similarity. Results from their experiments showed that chrF$_2$ correlates best out of the three evaluation metrics with human judgment for semantic similarity detection. We utilize this finding in later parts of our study and attempt to compare the outputs of our paraphrase detection models with respective chrF$_2$ scores.

Fadaee and Monz (2020): This research lays out a simple framework for constructing adversarial paraphrases through logical operations such as word insertion/deletion and numerical/gender substitution. The research correspondingly showed that such minor modifications in translation inputs could lead to disproportionately larger changes in translation outputs; thereby showing an adversarial effect. This research ultimately claimed that modern NMT models are unexpectedly *volatile*, or vulnerable, to adversarial attacks. We attempt to compare this claim with our findings in later parts of this study.

4 Experimental setup

4.1 Data sets

4.1.1 WMT19 en-de references and corresponding paraphrases

Freitag et al. (2020) builds on the premise that while automatic sequence evaluation metrics, such as BLEU, are important for NMT model evaluation; the presence of diverse translation references is also critical. Motivated by the observation that typical references show poor diversity, Freitag et al. (2020) focuses on two goals; namely creating additional high quality German WMT19 test references, as well as paraphrasing both existing (or legacy) and additional German WMT19 test references. These services were ultimately rendered by a professional translation service using different sets of linguists for different tasks to reduce systematic bias.

While these additional references serve the purpose of diversifying translation references, we see them as a source of high-quality semantically equivalent German paraphrases with varied lexical and syntactical features. Below are the key German data sets for that were used in our research.

WMT19 legacy test references: This refers to the existing `newstest2019` translation references with 1997 sentences. For abbreviation purposes, we refer to this data set as `WMT`.

WMT19 additional test references: This refers to additional references produced as a result of Freitag et al. (2020) with 1997 sentences. For abbreviation purposes, we refer to this data set as `AR`.

WMT19 legacy test paraphrased references: This refers to the paraphrased version of the existing `newstest2019` translation references produced as a result of Freitag et al. (2020) with 1997 sentences. For abbreviation purposes, we refer to this data set as `WMT.p`.

WMT19 additional test paraphrased references: This refers to the paraphrased version of the additional translation references produced as a result of Freitag et al. (2020) with 1997 sentences. For abbreviation purposes, we refer to this data set as `AR.p`.

For brevity, we concatenate the aforementioned data sets into WMT19 Legacy and WMT19 AR. Both data sets have 1997 pairs of semantically equivalent German paraphrases:

$$\text{WMT19 Legacy} = \{\text{WMT} \cup \text{WMT.p}\} \quad (6)$$

$$\text{WMT19 AR} = \{\text{AR} \cup \text{AR.p}\} \quad (7)$$

4.1.2 PAWS-X

PAWS-X is a cross-lingual adversarial data set for paraphrase identification released by Google Research (Yang et al., 2019). PAWS-X stems originally from the PAWS data set released by Zhang et al. (2019) which is an abbreviation for **P**araphrase **A**dversaries from **W**ord **S**crambling.

The original motivation behind the PAWS data set was that existing paraphrase detection data sets lacked non-paraphrase sentence pairs with high lexical overlap. The PAWS data set was therefore released to drive progress in creating models that utilize fine-grained structure and context of sentence pairs (Zhang et al., 2019).

The PAWS data set contains 108,463 paraphrase and non-paraphrase sentence pairs with high lexical overlap. These sentence pairs were bulk sourced from Wikipedia and Quora Question Pairs; followed by controlled word swapping and back translation to create challenging sentence pairs for paraphrase detection. The generated sentence pairs were finally evaluated for fluency and general quality by human raters (Zhang et al., 2019).

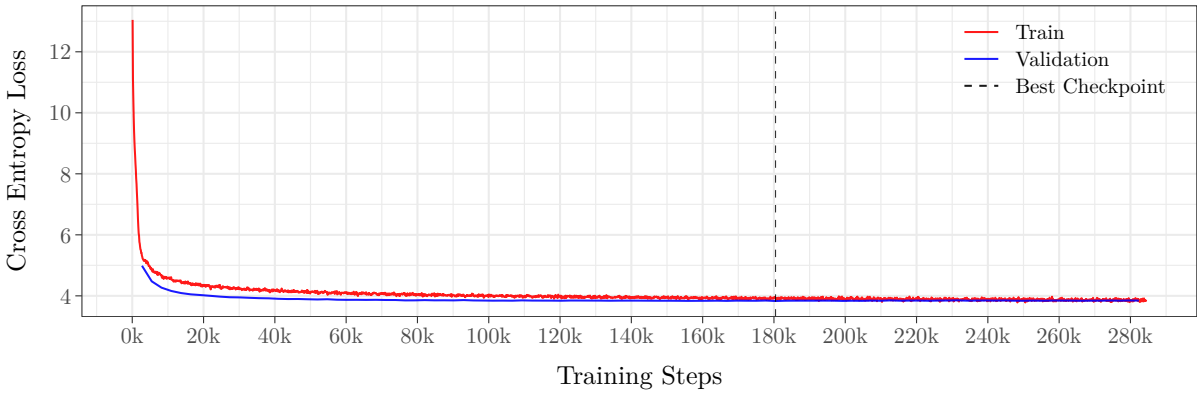

Figure 2: Training and validation cross entropy loss against training steps for Scaling NMT WMT16 Transformer

As noted in Yang et al. (2019), one limitation of adversarially generated data sets such as PAWS is their pre-dominant focus on the English language. In order to address this issue, Yang et al. (2019) released PAWS-X; which consists of 23,659 human translated evaluation sentence pairs and 296,406 machine-translated training sentence pairs whose source sentences were derived from the Wikipedia subset of the original PAWS data set. These sentence pairs were translated from English to six typologically distinct languages; namely French, Spanish, German, Chinese, Japanese and Korean.

The release of PAWS-X provides many advantages to the field of NLP, particularly the creation of a new benchmark to promote research in multilingual and zero-shot paraphrase detection. This can already be seen by the incorporation of PAWS-X into Google's recent Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark system (Hu et al., 2020).

4.1.3 WMT16 de-en

In this study, we replicate a non-SOTA NMT model from scratch based off the Scaling NMT WMT16 workflow (Ott et al., 2018). While the original implementation in Ott et al. (2018) is based on the `en→de` translation direction, our implementation trains a NMT model in the reverse translation direction; specifically `de→en`.

For this, we use WMT16 `de→en` training data with 4.5M sentence pairs, `newstest2013` as our validation set and `newstest2014` as our test set. We utilize a vocabulary of 32K symbols based off a joint source and target byte-pair encoding (BPE; Sennrich et al. 2015).

4.2 Models

4.2.1 FAIR WMT19 Transformer

We utilize FAIR's winning WMT19 single Transformer model as our SOTA NMT model. Focusing particularly on the `de→en` translation direction, the FAIR WMT19 Transformer was the top performing model in WMT19 with a SacreBLEU (Post, 2018) score of 40.8.

As per Ng et al. (2019), the key factors that led to SOTA performance include `langid` filtering of crawled bitext data, large-scale back translation as a form of data augmentation and noisy channel model reranking. We utilized this model directly from `fairseq` (Ott et al., 2019).

4.2.2 Scaling NMT WMT16 Transformer

We replicate the Scaling NMT WMT16 Transformer based on Ott et al. (2018) by training it from scratch. However, we reverse the original translation direction from `en→de` to `de→en`; such that we can ultimately use this model to translate WMT19 paraphrases from `de→en`. We intentionally choose this workflow since it would produce a non-SOTA transformer which would be useful for us downstream to introduce performance-dependent variance in the translation of WMT19 paraphrases.

Besides the aforementioned modification, we follow the same setup as per Ott et al. (2018). Specifically, we use a "big" transformer model based off Vaswani et al. (2017); with 6 blocks in the encoder and decoder networks. This model has a total of 210M parameters.

During training, we apply dropout (Srivastava et al., 2014) with probability 0.3 and utilize the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We use a learning rate

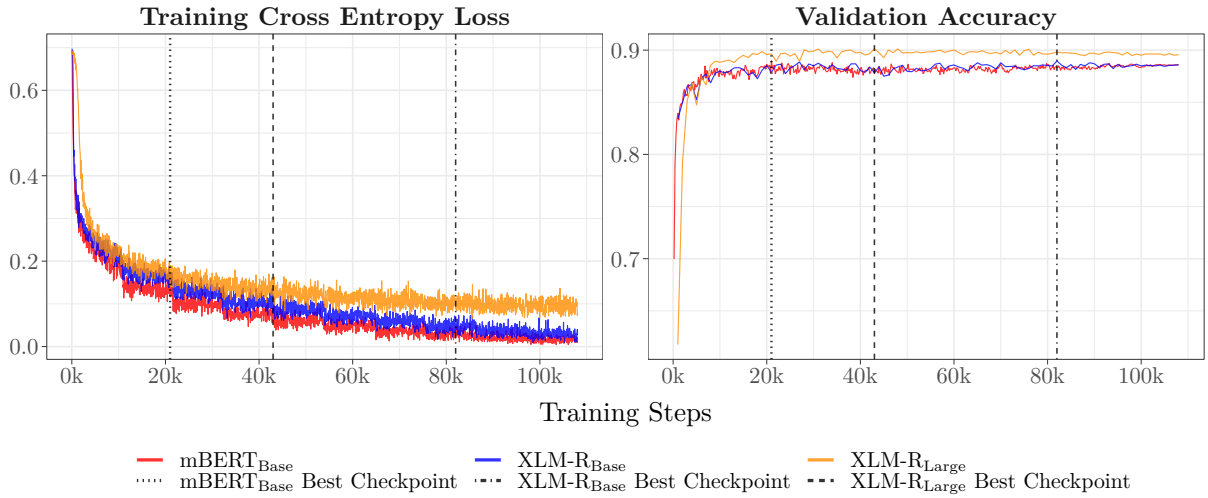

Figure 3: Training cross-entropy loss and validation accuracy against training steps for paraphrase detection models

schedule where the learning rate increases linearly for 4,000 steps from 1e-7 until 1e-3. The learning rate then decays proportionally to the inverse square root of the number of training steps. We utilize label smoothing with weight 0.1 for the uniform prior distribution over the vocabulary (Pereyra et al., 2017). We use large batch sizes with the maximum number of tokens per batch being 7168. Furthermore, we apply gradient accumulation for 8 steps before updating the model. We also exploit `fairseq`'s half precision floating point (FP16) functionality for more efficient training.

Finally, we train this model for 6 days on a single NVIDIA Tesla V100 16GB GPU. During training, we monitor the validation loss and enable checkpoint-saving for the best performing checkpoint on the validation set. We train the model up until ∼285K updates.

Our best performing checkpoint was saved at ∼180K updates as seen in Figure 2. For evaluation on the test set, we utilize beam search with a beam width of 5. Our final Scaling NMT WMT16 Transformer achieved a SacreBLEU (Post, 2018) score[2] of 31.0 on the `newstest2014` test set.

4.2.3 Paraphrase detection models

As noted in equation 4, paraphrase detection models are useful in computing probabilistic binary semantic equivalence spaces, or otherwise the S_L function. We follow a similar framework as per Google's XTREME benchmark (Hu et al., 2020) and fine-tune pre-trained multilingual transformer language models on the PAWS-X paraphrase detection task. We focus specifically on three multilingual transformer language models, specifically mBERT$_{Base}$ (104 languages; 172M parameters; Devlin et al. 2019), XLM-R$_{Base}$ (100 languages; 270M parameters; Conneau et al. 2019) and XLM-R$_{Large}$ (100 languages; 550M parameters; Conneau et al. 2019) using HuggingFace's `transformers` library (Wolf et al., 2019) with model variants optimized for sequence classification.

Language	mBERT$_B$	XLM-R$_B$	XLM-R$_L$
en	0.940	0.946	0.960
de	0.898	0.900	0.912
es	0.908	0.922	0.928
fr	0.922	0.917	0.933
ja	0.836	0.836	0.859
ko	0.841	0.847	0.870
zh	0.854	0.861	0.876
μ	0.886	0.890	**0.906**

Table 1: Language-specific summary of macro-F$_1$ scores of paraphrase detection models on the PAWS-X test set; languages are abbreviated based on ISO 639-1; B and L refer to base and large respectively

While our implementation is similar to that of Google's XTREME benchmark, we modify some aspects of the workflow to suit our needs. Most importantly, we fine-tune our multilingual language models on PAWS-X training data from all 7 languages instead of only English in order to reap the benefits of diverse multilingual data.

For all models, we enforce a maximum sequence length of 128 tokens since PAWS-X sentence pairs

[2]SacreBLEU signature: BLEU+case.mixed+lang.de-en+numrefs.1+smooth.exp+test.wmt14/full+tok.13a+version.1.4.12

generally fit it into this range. We use a modified Adam optimizer with decoupled weight decay regularization (Loshchilov and Hutter, 2017) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We train all models for 10 epochs or ∼110K updates with a global batch size of 32. We also use a linearly decaying learning rate schedule without warmup steps. Lastly, we monitor accuracy on the PAWS-X validation set for all languages in order to determine the best performing checkpoint.

Specific to $\text{mBERT}_{\text{Base}}$ and $\text{XLM-R}_{\text{Base}}$, we use a batch size of 32 without gradient accumulation and an initial learning rate of 2e-5. As for $\text{XLM-R}_{\text{Large}}$, we use an initial learning rate of 1e-6 and local batch size of 8 with 4 gradient accumulation steps to curb GPU out-of-memory (OOM) issues.

We fine-tune $\text{mBERT}_{\text{Base}}$, $\text{XLM-R}_{\text{Base}}$ and $\text{XLM-R}_{\text{Large}}$ for 14 hours, 15 hours and 2.5 days on a single NVIDIA GeForce GTX 1080 Ti 12GB GPU respectively. The best checkpoints are achieved and saved at ∼20K, ∼80K and ∼40K updates respectively, as seen in Figure 3.

As seen in Table 1, all three models perform well; especially on English and German. Overall, the best performing model on the PAWS-X test set is $\text{XLM-R}_{\text{Large}}$ with a macro-F_1 of 0.906.

4.3 Evaluation protocols

Given our German WMT19 Legacy and WMT19 AR paraphrase data sets defined in equations 6 and 7, we translate all pairs of paraphrases using both the FAIR WMT19 and the Scaling NMT WMT16 Transformers in the `de→en` translation direction using a beam width of 5. With this, we utilize the source German paraphrases along with their target-side English translations for further investigation.

4.3.1 Isometry on binary semantic equivalence spaces

Vector representation: We modify our representation of the S_L relations in equation 5 in order to have a more concise vectorized form of the relationship. We assign the probability threshold ϵ from equation 4 a constant value of 0.5 for all computations of S_L in this research:

$$\mathbf{S_{XY}} = \begin{bmatrix} S_X(s_1, s_2) \\ S_Y(f(s_1), f(s_2)) \end{bmatrix} \tag{8}$$

$$\mathbf{S_{XY}^\intercal} = \begin{bmatrix} S_X(s_1, s_2) & S_Y(f(s_1), f(s_2)) \end{bmatrix} \tag{9}$$

Multi-model decisions: Given this formulation of $\mathbf{S_{XY}^\intercal}$, it is worth noting that each of the three paraphrase detection models would compute a separate $\mathbf{S_{XY}^\intercal}$ term. Since our paraphrase detection models were evaluated to have non-zero (albeit small and similar) error rates, we decide to use the $\mathbf{S_{XY}^\intercal}$ outputs of all three models and compute the statistical mode of the three outputs. Such a majority decision would provide more confidence in any particular decision from the models. For simplicity, we assign the statistical mode function with the symbol $\mathbf{M}$. We therefore compute $\mathbf{M}(\mathbf{S_{XY}^\intercal})$ to check for a majority decision over the three paraphrase detection models. We assign the empty set $\emptyset$ in case no majority decision exists.

Discrete possibilities: Since the output of S_L falls in the binary set of $\{0, 1\}$, we can effectively compute all five possibilities of the consequent majority decision $\mathbf{M}(\mathbf{S_{XY}^\intercal})$. We describe these in the context of binary semantic equivalence spaces.

1. $\mathbf{M}(\mathbf{S_{XY}^\intercal}) = [1, 1]$: Both the source and target pairs of sentences were evaluated to be paraphrases and are therefore semantically equivalent on both sides. This qualifies as *isometric behaviour*.

2. $\mathbf{M}(\mathbf{S_{XY}^\intercal}) = [0, 0]$: Both the source and target pairs of sentences were evaluated to be non-paraphrases and are therefore not semantically equivalent on both sides. While this does qualify as approximate isometry according to equation 5, we also observe this scenario is less conclusive for determining isometry compared to $\mathbf{M}(\mathbf{S_{XY}^\intercal}) = [1, 1]$ because of a looser bound associated with $S_L = 0$ as per equation 3. We therefore consider this behaviour to be *ambiguous*.

3. $\mathbf{M}(\mathbf{S_{XY}^\intercal}) = [0, 1]$: The source sentences are evaluated to be non-paraphrases while the target sentences are evaluated to be paraphrases. This implies that translation resulted in the sentence pair becoming semantically equivalent while they were not before. This qualifies as anisometric behaviour and could be an interesting scenario to investigate further. We define this as *type-1 anisometric behaviour*.

4. $\mathbf{M}(\mathbf{S_{XY}^\intercal}) = [1, 0]$: The source sentences are evaluated to be paraphrases while the target sentences are evaluated to be non-paraphrases.

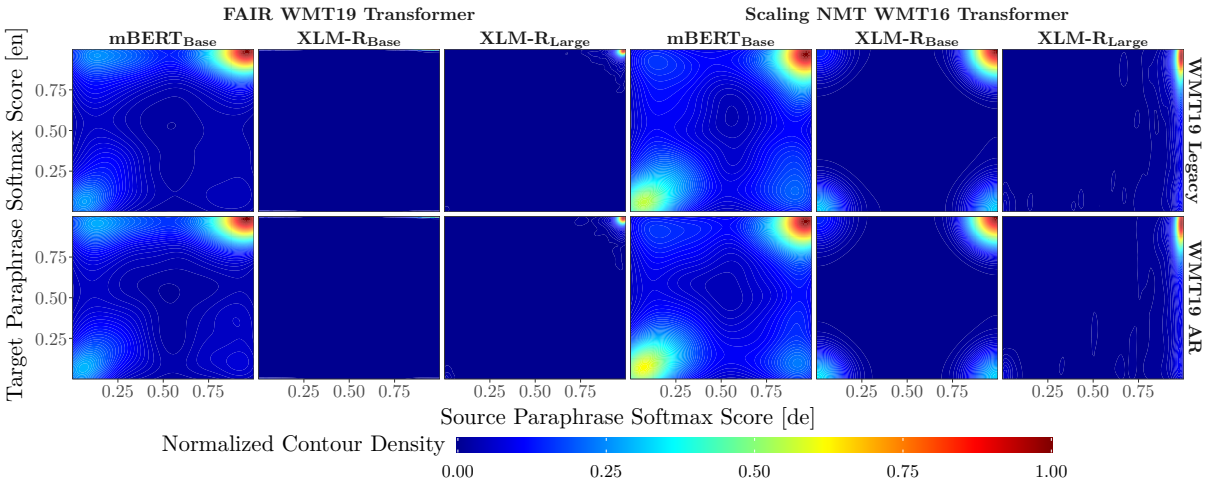

Figure 4: Normalized contour densities for target paraphrase softmax score against source paraphrase softmax score by NMT models, paraphrase detection models and input data sets

This implies that translation resulted in the sentence pair becoming semantically inequivalent while they were not before. This qualifies as anisometric behaviour and could imply weak performance of the model on one of the sentences; or possibly some adversarial paraphrasing. We define this as *type-2 anisometric behaviour*.

5. $\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}}) = \emptyset$: This implies that all models had different decisions and therefore no majority decision was reached. We consider this behaviour to be *ambiguous*.

Frequency analysis: Given the five aforementioned possibilities of $\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}})$, we measure the frequency of each possibility for each model. This give us an insight into the semantically isometric behaviour of the models and would allow us to conduct comparisons across them.

Sentence-level analysis: In order to check the veracity of the $\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}})$ outputs, we analyze selected individual sentence pairs and their target translations, and attempt to interpret these results.

4.3.2 Relationship between chrF$_2$ and semantic equivalence

We return to one of the observations from Michel et al. (2019), specifically that the chrF$_2$ automatic sequence evaluation metric (Popović, 2015) tends to correlate most positively with human judgment of semantic similarity compared to BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014). We use our current research to further investigate the relationship between semantic equivalence and the chrF$_2$ automatic sequence evaluation metric. We replicate the chrF$_2$ setup from Michel et al. (2019) by using the default `sacrebleu` implementation (Post, 2018) with a n-gram upper limit of 6 and β value of 2

Non-commutativity of chrF$_2$: While replicating the chrF$_2$ setup of Michel et al. (2019), we observe that this particular formulation with $\beta = 2$ results in the chrF$_2$ metric being non-commutative, where s_1 and s_2 are input sentences:

$$\text{chrF}_2(s_1, s_2) \neq \text{chrF}_2(s_2, s_1) \tag{10}$$

Non-commutativity is an emergent property of chrF$_2$ with $\beta = 2$ since this would assign two times more weight to recall than precision (Popović, 2015); which places an internal bias on input order. While this non-commutative formulation of chrF$_2$ would be useful for evaluating NMT models with explicit hypotheses and references, this would not be optimal as a semantic similarity metric since one would expect a semantic similarity metric to be commutative and unbiased towards input order.

Commutative variant of chrF$_2$: As an alternative, we simulate commutativity by averaging the chrF$_2$ values for both input orders and introduce $\overline{\text{chrF}_2}$ as a commutative variant for chrF$_2$:

$$\overline{\text{chrF}_2}(s_1, s_2) = \frac{\text{chrF}_2(s_1, s_2) + \text{chrF}_2(s_2, s_1)}{2} \tag{11}$$

$$\therefore \quad \overline{\text{chrF}_2}(s_1, s_2) = \overline{\text{chrF}_2}(s_2, s_1) \tag{12}$$

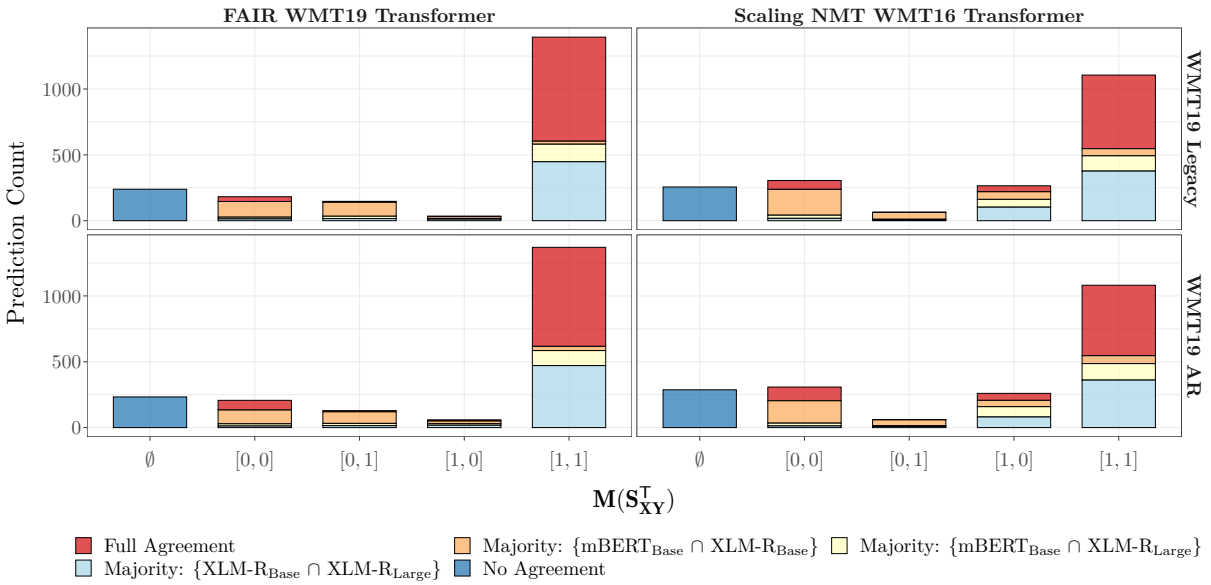

Figure 5: Frequency distribution of $\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}})$ by NMT models and input data sets; filling colors indicate finer details on the type of majority decision

$\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}})$	FAIR WMT19 Transformer			Scaling NMT WMT16 Transformer		
	WMT19 Legacy	WMT19 AR	μ	WMT19 Legacy	WMT19 AR	μ
$[1, 1]$	**0.698**	**0.686**	**0.692**	0.554	0.541	0.548
$[0, 0]$	0.091	0.104	0.097	0.153	0.154	0.153
$[0, 1]$	**0.074**	**0.065**	**0.069**	0.033	0.031	0.032
$[1, 0]$	0.018	0.030	0.024	**0.133**	**0.130**	**0.131**
$\emptyset$	0.120	0.117	0.118	0.128	0.144	0.136

Table 2: Relative frequency distribution for $\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}})$ by NMT models and input data sets; μ indicates the macro-average of the relative frequency over the input data sets for each model

We see this as a more optimal alternative than changing β to 1 since this might veer further away from the experimental setup of Michel et al. (2019). Such commutative variants of automatic sequence evaluation metrics have also been considered for BLEU and METEOR in Wieting et al. (2019) and were termed *symmetric* instead of commutative.

Mapping $\overline{\text{chrF}_2}$ to S_L: We refer back to the majority decisions of the paraphrase detection models and remove all sentence pairs where $\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}}) = \emptyset$. We assume all remaining sentence pairs have been *confidently* tagged by the paraphrase detection models. We assign the remaining sentence pairs with their respective $\overline{\text{chrF}_2}$ and S_L values for the source and target sides.

Correlation between $\overline{\text{chrF}_2}$ and S_L: Finally, we replicate a similar statistical procedure as per Michel et al. (2019) and compute the Pearson correlation coefficient r_{xy} and a corresponding statistical t-test. For the t-test, we set the null hypothesis H_0 to be that there exists a non-positive correlation between $\overline{\text{chrF}_2}$ and S_L; while the alternative hypothesis H_1 is that there exists a positive correlation between $\overline{\text{chrF}_2}$ and S_L. We interpret the strength of correlations using guidelines from Schober et al. (2018).

5 Results

5.1 Isometry on binary semantic equivalence spaces

5.1.1 Paraphrase detection softmax scores

Figure 4 shows a normalized contour density estimate for paraphrase detection softmax scores. These scores are grouped by NMT models, input data sets and paraphrase detection models. We observe that $\text{mBERT}_{\text{Base}}$ generally shows more variance in softmax scores compared to the XLM-R

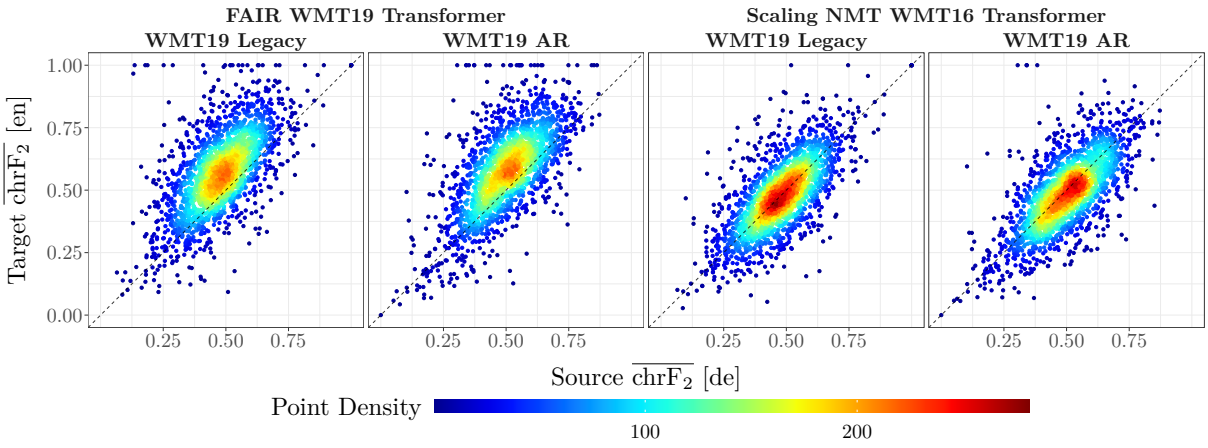

Figure 6: Distributions of target-side $\overline{\text{chrF}_2}$ against source-side $\overline{\text{chrF}_2}$ by NMT models and input data sets

models. We can also observe that all models show more variance for translation outputs from the Scaling NMT WMT16 Transformer compared to those from the FAIR WMT19 Transformer. The softmax distributions are generally similar between the WMT19 Legacy and WMT19 AR input data sets.

5.1.2 Frequency analysis

Figure 5 shows a visual breakdown of absolute frequencies of $\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}})$ by NMT models and input data sets. Table 2 shows a tabular breakdown of relative frequencies of $\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}})$ by NMT models and input data sets. Below are the key observations in the context of binary semantic equivalence spaces.

Isometry: We observe a higher proportion of isometric behaviour with the FAIR WMT19 Transformer (69.2%) compared to the Scaling NMT WMT16 Transformer (54.8%).

Type-1 anisometry: We observe a higher proportion of type-1 anisometric behaviour with the FAIR WMT19 Transformer (6.9%) compared to the Scaling NMT WMT16 Transformer (3.2%).

Type-2 anisometry: We observe a lower proportion of type-2 anisometric behaviour with the FAIR WMT19 Transformer (2.4%) compared to the Scaling NMT WMT16 Transformer (13.1%).

Ambiguity: We observe a lower proportion of ambiguous samples with the FAIR WMT19 Transformer (21.5%) compared to the Scaling NMT WMT16 Transformer (28.9%).

Model agreement: Based on Figure 5, we observe that the majority of agreements are full agreements, followed by XLM-R_{Base} and XLM-R_{Large} agreements, $\text{mBERT}_{\text{Base}}$ and XLM-R_{Base} agreements and finally $\text{mBERT}_{\text{Base}}$ and XLM-R_{Large} agreements.

5.2 Correlation between $\overline{\text{chrF}_2}$ and S_L

5.2.1 Source and target $\overline{\text{chrF}_2}$ distributions

Figure 6 shows the distribution of $\overline{\text{chrF}_2}$ over the source and target sides, grouped over NMT models and input data sets. We observe a larger variance of $\overline{\text{chrF}_2}$ points for the FAIR WMT19 Transformer outputs compared to those from the Scaling NMT WMT16 Transformer.

Furthermore, we can observe a larger mean $\overline{\text{chrF}_2}$ value for the FAIR WMT19 Transformer compared to the Scaling NMT WMT16 Transformer. This can be inferred from the general distribution of points from the former being above the diagonal compared to those from the latter being centered near the diagonal. We can also observe more cases where $\overline{\text{chrF}_2} = 1$ on the target side but $\overline{\text{chrF}_2} \neq 1$ on the source side for the FAIR WMT19 Transformer compared to the Scaling NMT WMT16 Transformer.

5.2.2 Correlation analysis

Figure 7 shows the distribution of $\overline{\text{chrF}_2}$ against S_L by over NMT models, input data sets and source-target origins. Table 3 shows a breakdown of Pearson correlation coefficients r_{xy} for $\overline{\text{chrF}_2}$ and S_L and results of the statistical t-test by NMT models and input data sets.

Overall, we observe a significant positive correlation between $\overline{\text{chrF}_2}$ and S_L with mean Pearson correlation coefficient r_{xy} values of 0.256 and 0.250 for the FAIR WMT19 Transformer and the Scaling NMT WMT16 Transformer respectively.

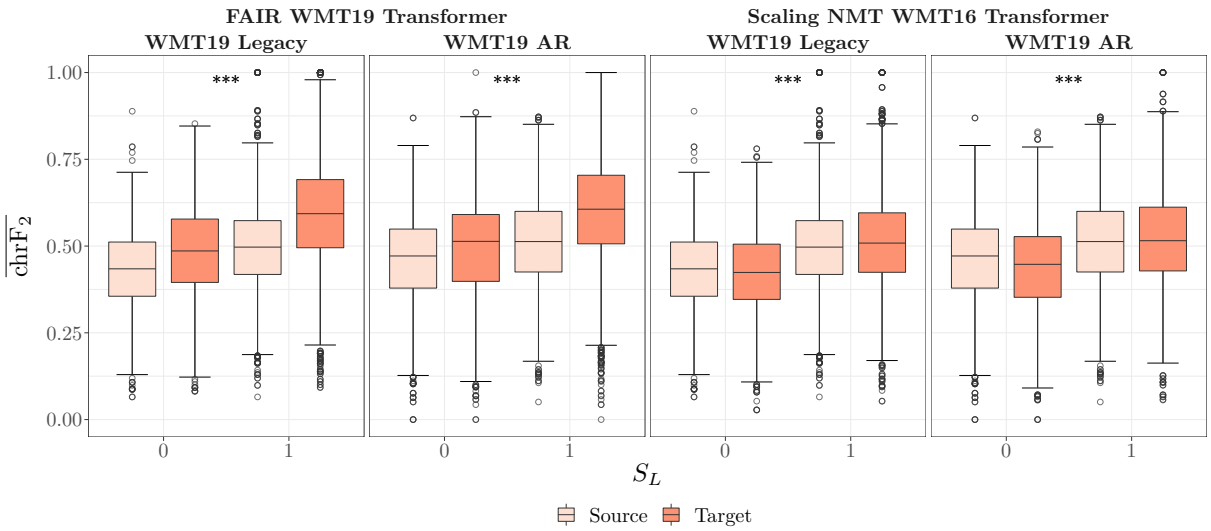

Figure 7: Distributions of $\overline{\text{chrF}_2}$ against S_L by NMT models and input data sets ; *** indicates a statistically significant positive correlation between $\overline{\text{chrF}_2}$ and S_L with $p \leq 0.001$ for the one-tailed t-test

Statistic	FAIR WMT19 Transformer			Scaling NMT WMT16 Transformer		
	WMT19 Legacy	WMT19 AR	μ	WMT19 Legacy	WMT19 AR	μ
r_{xy}	0.269	0.243	**0.256**	0.269	0.231	**0.250**
$H_1 : r > 0$	***	***	—	***	***	—
Correlation	Weak	Weak	—	Weak	Weak	—

Table 3: Tabular summary of Pearson correlation coefficients r_{xy}, t-test alternative hypothesis and correlation strength interpretation (Schober et al., 2018) by NMT models and input data sets; *** indicates $p \leq 0.001$ for the one-tailed t-test

The one-tailed t-test used to ascertain significance showed a strongly significant positive correlation with $p \leq 0.001$. According to the interpretation guidelines from Schober et al. (2018), this range of r_{xy} would imply a weak correlation strength between $\overline{\text{chrF}_2}$ and S_L.

6 Discussion

6.1 Isometry on binary semantic equivalence spaces

6.1.1 Isometry and robustness to paraphrases

We observe that the SOTA FAIR WMT19 Transformer exhibits more frequent semantically isometric behaviour (69.2%) on binary semantic equivalence spaces compared to the non-SOTA Scaling NMT WMT16 Transformer (54.8%). This confirms our initial hypothesis, albeit with a small sample size of two models, that the frequency of semantically isometric behaviour correlates positively with general model performance.

From our perspective, this could imply that the FAIR WMT19 Transformer is more robust to lexically and syntactically diverse paraphrases; in the sense that it more often correctly and consistently transfers the original semantics of paraphrases compared to the Scaling NMT WMT16 Transformer. We predict that this robustness to semantically-equivalent lexical and syntactical variations arose largely because of heavy data augmentation from large-scale back translation, which ultimately introduced lexical and syntactical variations into the training data of the FAIR WMT19 Transformer. Large-scale back translation is purportedly also the process that led to the largest improvement in translation quality for the FAIR WMT19 Transformer's `de→en` translation task (Ng et al., 2019).

An effective way of testing the veracity of the aforementioned prediction would be to train the FAIR WMT19 Transformer without back translation and observe the changes to the frequency of semantically isometric behaviour. We were unable

$\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}})$	Type	Sentence in WMT19 AR	Paraphrase in WMT19 AR
$[1, 1]$	Source	Glocken von St. Martin verstummen, da Kirchen in Harlem kämpfen	St. Martin‘s Glocken läuten nicht mehr, weil in Harlem ein Kampf der Kirchen in Gang ist
	Target	Bells of St. Martin fall silent as churches struggle in Harlem	St Martin’s bells no longer ring as churches battle it out in Harlem
$[0, 0]$	Source	Davor empfangen die Rangers am Donnerstag in der Europa League Rapid Wien.	Rapid Wien steht am Donnerstag im Rahmen der Europa League den Rangers gegenüber.
	Target	Rangers host Rapid Vienna in the Europa League on Thursday.	Rapid Vienna face Rangers in the Europa League on Thursday.
$[0, 1]$	Source	Das Tor fiel in der 29. Minute.	In der 29. Minute kam es zum Tor.
	Target	The goal came in the 29th minute.	The goal came in the 29th minute.
$[1, 0]$	Source	Laut Berichten in Lokalmedien wurde auf einem Markt im Südwesten von China von einem Schwein angegriffen und getötet.	Aus Meldungen der lokalen Medien geht hervor, dass im Südwesten Chinas ein Mann auf einem Markt durch ein Schwein zu Tode gebissen wurde.
	Target	According to reports in local media, a pig was attacked and killed in a market in southwest China.	Local media reported that a man was bitten to death by a pig in a market in southwest China.

Table 4: Examples of German sentence pairs and their English translations corresponding to the meaningful values of $\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}})$ which had full agreement across the three paraphrase detection models; translations were derived from the FAIR WMT19 Transformer on the WMT19 AR German paraphrase input data set; top-down indices of the above sentence pairs are 22, 307, 527 and 178 respectively

to conduct such a comparison and deem this as a possible subject of further study.

6.1.2 Sentence-level analysis

In order to gain ground-level insight of the translation process and associated isometry on binary semantic equivalence spaces, we analyze selected sentences individually. For brevity, we only analyze the outputs of the FAIR WMT19 Transformer on the WMT19 AR data set.

Table 4 shows a breakdown of sentence pairs from four meaningful $\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}})$ categories which were predicted with full agreement from all three paraphrase detection models. We provide manual analyses of sentence pairs and their translations, and attempt to interpret the majority decisions made by the paraphrase detection models.

$\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}}) = [1, 1]$**:** From manual assessment, both sentence pairs on the source and target side are grammatically well-formed and have the same meaning. They are correspondingly classified as paraphrases on both sides.

$\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}}) = [0, 0]$**:** From manual assessment, we observe that the source sentences are non-paraphrases because the sentence in `AR` has a critical time marker “*davor*” which is not present in the paraphrase. Furthermore without sport-specific context, the verbs “*empfangen*” and “*gegenüberstehen*” would be interpreted as semantically inequivalent. This is also the case for the target sentences, where “*empfangen*” was translated to be the verb “host” while “*gegenüberstehen*” was translated to “face”. Overall without sport-specific context, we would argue that the source sentences were not well paraphrased and resulted in the non-paraphrase classification. The target sentences were translated well, but still retained the semantic difference from the source sentences; resulting in them being classified as non-paraphrases.

$\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}}) = [0, 1]$**:** From manual assessment and without sport-specific context, we observe that the source sentences are non-paraphrases because the verbs "*fallen*" and "*kommen*" would be interpreted as semantically inequivalent; possibly leading to the non-paraphrase classification. Based on the target sentences, it appears the NMT model captured sport-specific context and translated both source sentences to the same target sentence. These were naturally classified as paraphrases on the target side.

We interpret the behaviour observed here as *context regularization* from the NMT model, where two lexically and syntactically diverse paraphrases were translated to the same exact target sentence because the NMT model incorporated the relevant context into its translation.

$\mathbf{M}(\mathbf{S}^{\mathsf{T}}_{\mathbf{XY}}) = [1, 0]$**:** From manual assessment, we observe that the source sentence in AR was grammatically ill-formed as the passive subject of the verbs "*angreifen*" and "*töten*" was missing. The aforementioned sentence's paraphrase was however grammatically well-formed. It appears that the paraphrase detection models did not penalize the lack of a passive subject in the first sentence and still classified the source sentences as paraphrases. On the target side, the NMT model translated the paraphrases such that they had opposing semantic roles; where "a pig was attacked and killed" in the first sentence while "a man was bitten to death by a pig" in the paraphrase. This resulted in them being classified as non-paraphrases.

We interpret that the NMT model mistranslated the source AR sentence and lost part of its original semantics. Despite the absence of a passive subject, "*von einem Schwein*" clearly assigns the pig as the passive object. This was however not reflected in the translation, where the pig became the passive subject.

6.2 Correlation between $\overline{\text{chrF}_2}$ and S_L

Our results show that $\overline{\text{chrF}_2}$ and S_L are weakly but significantly positively correlated. In general, this would support the claim from Michel et al. (2019) that chrF_2 and semantic similarity are positively correlated. However, the Pearson correlation coefficients observed in our results are roughly half in size compared to those observed in Michel et al. (2019) which had a range of ∼0.5-0.6.

Our smaller correlation coefficients could be attributed to the lack of granularity in our semantic equivalence measures; which were binary in our case but were distributed into 5 granular ordinal categories in Michel et al. (2019).

Michel et al. (2019) further compared the correlation coefficients of chrF_2 with those from the BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014) automatic sequence evaluation metrics. We were unable to conduct such comparisons in our study and deem this as a possible subject of further study.

6.3 Volatility of NMT models to paraphrases

Fadaee and Monz (2020) claimed that NMT models are unexpectedly volatile to adversarial paraphrases created using logical operations such as word insertion/deletion and numerical/gender substitution. Their study showed changes in translation quality for 26% and 19% of sentence variations for their RNN and Transformer models respectively.

While we did not conduct similar adversarial paraphrasing in our study, we still investigated the performance of NMT models when translating sentences that are non-trivial paraphrases of one another. Given the observations of Fadaee and Monz (2020) on NMT model volatility, we would have expected the relative frequency of type-2 anisometry to also be in the range of ∼15-30% for both our models.

Instead, we observed a relative frequency of type-2 anisometry to be 2.4% and 13.1% for the FAIR WMT19 and Scaling NMT WMT16 Transformers respectively. These relative frequencies are surprisingly low, even for the latter non-SOTA model. The differences in the "rate of volatility" can be attributed to many factors. In support of Fadaee and Monz (2020), our paraphrases originating from Freitag et al. (2020) were not targeted for adversarial purposes. It is therefore expected that the adversarial effect of such paraphrases would be lower than the paraphrases constructed by Fadaee and Monz (2020) which were designed with an adversarial goal.

On the other hand, Fadaee and Monz (2020) used non-SOTA RNN and Transformer models in their experiments and the volatility observed could have been a result of these less performant models. It would be beneficial if Fadaee and Monz (2020) would re-run their experiments estimating volatility on SOTA NMT models, such as the FAIR WMT19 Transformer. This could elucidate whether the volatility observed is endemic to all NMT mod-

els or just to non-SOTA models.

7 Conclusions

Our research investigates the isometric behaviour of NMT models on binary semantic equivalence spaces. By using two NMT models of varying performance, we were able to confirm our initial hypothesis that the frequency of semantically isometric behaviour in NMT models correlates positively with general model performance, albeit with a small sample size of two models.

Next, we provide evidence to support the claim of Michel et al. (2019) that chrF_2 is significantly positively correlated with semantic similarity. Our experiments however show correlation coefficients which are roughly half in size compared to those reported in Michel et al. (2019).

Finally, we provide light counter-evidence to the claim in Fadaee and Monz (2020) that NMT models exhibit high rates of volatile behaviour ($\sim$19-26%) when provided paraphrased input sentences. While our input paraphrases were not adversarial in nature, they were still lexically and syntactically diverse and showed considerably smaller rates of volatile behaviour ($\sim$2-13%) when translated with our NMT models. We suspect that the high rate of volatility observed by Fadaee and Monz (2020) could be partially attributed to the utility of non-SOTA NMT models in their experiments. We would therefore recommend re-running the experiments with SOTA NMT models, such as the FAIR WMT19 Transformer.

8 Further work

In this study, we compared the isometric behaviour of the SOTA FAIR WMT19 Transformer against the non-SOTA Scaling NMT WMT16 Transformer on binary semantic equivalence spaces. We would recommend comparing this isometric behaviour of the SOTA model with another better performing non-SOTA model; such as a freshly trained FAIR WMT19 Transformer without large-scale back translation. This might help to narrow down the main cause(s) of the more frequent semantically isometric behaviour in the SOTA NMT model.

While we made a case for utilizing commutative semantic similarity metrics in our research, one major logical limitation of our current paraphrase detection models is that they are themselves non-commutative; since they were trained with a specific sentence input order. We could recommend further research on commutative paraphrase detection models, such as those which compare cosine similarities of sentence embeddings.

Finally, we recommend computing the correlation coefficients of commutative BLEU scores on our sentence pairs against S_L. This might help to elucidate whether BLEU scores have a stronger or weaker correlation to semantic similarity compared to chrF_2, which could provide additional evidence for or against the findings of Michel et al. (2019).

References

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Harold Scott Macdonald Coxeter. 1961. Introduction to geometry.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

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

Marzieh Fadaee and Christof Monz. 2020. The unreasonable volatility of neural machine translation models. *arXiv preprint arXiv:2005.12398*.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. *ArXiv*, abs/2004.06063.

C. Hegde, A. C. Sankaranarayanan, W. Yin, and R. G. Baraniuk. 2015. Numax: A convex approach for learning near-isometric linear embeddings. *IEEE Transactions on Signal Processing*, 63(22):6109–6121.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task

benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Paul Michel, Xian Li, Graham Neubig, and Juan Miguel Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. *arXiv preprint arXiv:1903.06620*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. *arXiv preprint arXiv:1907.06616*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

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

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

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

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

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 *Advances in neural information processing systems*, pages 5998–6008.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. In *Proc. of EMNLP*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.