Investigating the isometric properties 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 hypothesize that well-performing Neural Machine Translation (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. We begin by utilizing two NMT models of varying performance to translate semantically-equivalent paraphrases based off WMT19 test data references. In order to quantify and simplify the notion of a semantic metric space, we treat it as a probabilistic binary semantic-equivalence space indicating either semantic equality or inequality; achieved by fine-tuning three transformer-based language models on Google's PAWS-X paraphrase detection task. By using the paraphrase detection outputs, we investigate the frequency and composition of semantically isometric behaviour in the NMT models' inputs and outputs.

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 hypothesize 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.

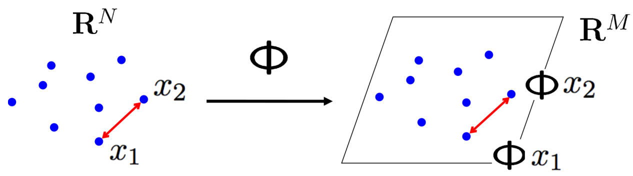

Figure 1: Illustration of isometry in higher dimensional functional transformations (Hegde et al., 2015)

In order to conduct our investigation, we start by acquiring semantically equivalent paraphrases of WMT19 legacy and additional test references from Freitag et al. (2020) for en→de. 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), in order to translate the aforementioned paraphrases in the de→en translation direction. We use the former model pre-trained from `fairseq` (Ott et al., 2019) and train the latter model from scratch.

Next, we utilize three well-performing paraphrase detection models to approximate isometry in the NMT models' translations. These paraphrase detection models are based off $\text{mBERT}_{\text{Base}}$ (Devlin et al., 2019), $\text{XLM-R}_{\text{Base}}$ 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 finally investigate the frequency and composition of semantically isometric behaviour in the NMT models' inputs and outputs.

2 Isometry and approximations

The concept of isometry in the context of 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 sentence length N, $f : \mathbb{R}^{V\times N} \rightarrow \mathbb{R}^{V'\times N'}$ refers to the NMT model's inference function 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:

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

While elegant, this representation of isometry and a semantic distance metric is problematic. Firstly, exact isometry might not be a practical condition to achieve given real-life data instances with stochastic noise. Next, 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).

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

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

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 equality ($S_L = 1$) or inequality ($S_L = 0$) 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 equality than from semantic inequality, due to the presence of a tighter bound for the former than the latter.

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:

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

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 targetted probe into semantic *anisometry* of NMT models, compared to our research which would be an untargetted 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. Additonally, this research compared three automatic sequence evaluation metrics, specifically `BLEU` (Papineni et al., 2002), `METEOR` (Denkowski and Lavie, 2014) and `chrF` (Popović, 2015), against human judgment for evaluating semantic similarity. Results from their experiments showed that `chrF` correlates best out of the three similarity 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` 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 could lead to disproportionately larger changes in translation outputs; thereby showing an adversarial effect. This research ultimately claimed that modern NMT models are generally *volatile*, or vulnerable, to targetted adversarial attacks. We attempt to compare this claim with our findings in later parts of this study.

4 Experimental setup

4.1 Data sets

Below we describe the key data sets that we were used in our research.

4.1.1 WMT19 en-de references and corresponding paraphrases

Freitag et al. (2020) builds on the premise that while automatic 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 WMT19 test references, as well as paraphrasing both existing (or legacy) and additional WMT19 test references in the `en`$\rightarrow$`de` translation direction. These services were ultimately rendered by a professional translation service using different sets of linguists for different tasks to reduce systematic bias. Below are the key resulting data sets for `en`$\rightarrow$`de` that were used in our research.

WMT19 legacy test references: This refers to the existing `newstest2019` translation references without any additional references or paraphrasing. For abbrevation purposes, we refer to this dataset as `WMT`.

WMT19 additional test references: This refers to additional references produced as a result of Freitag et al. (2020). For abbrevation purposes, we refer to this dataset 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). For abbrevation purposes, we refer to this dataset 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). For abbrevation purposes, we refer to this dataset as `AR.p`.

For brevity, we further simplify the aforementioned datasets as follows:

$$\text{WMT19 Legacy} = \{\text{WMT} \cup \text{WMT.p}\} \tag{6}$$

$$\text{WMT19 AR} = \{\text{AR} \cup \text{AR.p}\} \tag{7}$$

4.1.2 PAWS-X

PAWS-X is a cross-lingual adversarial dataset for paraphrase identification released by Google Research (Yang et al., 2019). PAWS-X stems originally from the PAWS dataset 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 dataset was that existing paraphrase detection datasets lacked non-paraphrase sentence pairs with high lexical overlap. The PAWS dataset was therefore released to drive progress in creating models that utilize fine-grained structure and context of sentence pairs.

The PAWS dataset 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.

As noted in Yang et al. (2019), one limitation of adversarially generated datasets 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 PAWS evaluation sentence pairs and 296,406 machine-translated training sentence pairs derived from the Wikipedia subset of the original PAWS dataset. 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

As mentioned previously, 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`$\rightarrow$`de` translation direction, our implementation trains a NMT model in the reverse translation direction; specifically `de`$\rightarrow$`en`.

For this, we use WMT16 `en`$\rightarrow$`de` training data with $\sim$4.5M sentence pairs and reverse it to the desired `de`$\rightarrow$`en` translation direction. Furthermore, we use `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).

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.

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.

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.

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

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*.