Analysing Translation Artifacts: A Comparative Study of LLMs, NMTs, and Human Translations

Fedor Sizov[1] **Cristina España-Bonet**[2] **Josef van Genabith**[1,2]
Roy Xie[3] **Koel Dutta Chowdhury**[1]
[1]Saarland University, Saarland Informatics Campus [2]DFKI GmbH [3]Duke University
{cristinae, josef.van_genabith}@dfki.de,
sife00002@stud.uni-saarland.de, ruoyu.xie@duke.edu, koeldc@lst.uni-saarland.de

Abstract

Translated texts exhibit a range of characteristics that make them appear distinct from texts originally written in the same target language. With the rise of Large Language Models (LLMs), which are designed for a wide range of language generation and understanding tasks, there has been significant interest in their application to Machine Translation. While several studies have focused on improving translation quality through fine-tuning or few-shot prompting techniques, there has been limited exploration of how LLM-generated translations qualitatively differ from those produced by Neural Machine Translation (NMT) models, and human translations. Our study employs explainability methods such as Leave-One-Out (LOO) and Integrated Gradients (IG) to analyze the lexical features distinguishing human translations from those produced by LLMs and NMT systems. Specifically, we apply a two-stage approach: first, classifying texts based on their origin —whether they are original or translations— and second, extracting significant lexical features (highly attributed input words) using post-hoc interpretability methods. Our analysis shows that different methods of feature extraction vary in their effectiveness, with LOO being generally better at pinpointing critical input words and IG capturing a broader range of important words. Finally, our results show that while LLMs and NMT systems can produce translations of a good quality, they still differ from texts originally written by native speakers. We find that while some LLMs more closely resemble human translations, traditional NMT systems show distinct differences, particularly in their use of linguistic features. [1]

1 Introduction

The rapid development of large language models (LLMs) (Radford et al., 2019; Raffel et al., 2020a; Touvron et al., 2023; Lu et al., 2024; Team et al., 2024a; Groeneveld et al., 2024; Alves et al., 2024) has significantly advanced natural language processing (NLP), also in the domain of Machine Translation (MT) (Zhang et al., 2023; Zhu et al., 2024) with studies covering various approaches such as document-level literary translation (Karpinska and Iyyer, 2023), paragraph-level post-editing with LLMs (Thai et al., 2022), sentence-level translation (Vilar et al., 2022; Jiao et al., 2023), examining hallucinations in LLM-generated translations (Guerreiro et al., 2023), and leveraging LLMs for evaluation (Kocmi and Federmann, 2023). These efforts reflect the ongoing shift toward exploring how well LLMs perform MT compared to traditional NMT systems.

Although previous work (Zhu et al., 2024; Vilar et al., 2022; Raunak et al., 2023) have explored how LLMs and traditional Neural Machine Translation (NMT) systems develop translation capabilities, as well as the qualitative differences in their outputs and the factors that impact their performance, a critical gap remains: the comparison of translations generated by LLMs and NMT models to those produced by human translators (HT) and texts originally written by native speakers in the target language. This comparison raises questions about translation divergence, as reflected in surface-level (structural) differences in translations arising from cross-linguistic variations or translator preferences (Luo et al., 2024).

Such divergences are well-documented in human translations (HT), where translators often make structural choices that vary significantly from the text originally written in the target language (Deng and Xue, 2017; Nikolaev et al., 2020). In contrast, traditional NMT outputs typically exhibit less diversity and more literal translations, lacking significant structural variation (Freitag et al., 2020; Bizzoni et al., 2020). Similarly, Vyas et al. (2018); Briakou and Carpuat (2020) focus on identifying

[1]We release our code publicly at https://github.com/SFB1102/B6-analysing-translation-artifacts

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 1183–1199
November 15–16, 2024.

semantic divergences in translations that are not fully equivalent to the original source texts. Recent findings, however, indicate that LLMs tend to produce translations that are less literal compared to NMT models (Vilar et al., 2022; Raunak et al., 2023), suggesting that LLMs may bridge the gap between the rigid literalness of NMT models and the flexibility of human translations. Understanding these divergences is crucial for advancing translation technologies and ensuring their responsible and effective use. Specifically, this leads us to investigate the following research questions: **how do LLMs, NMT models, and HT outputs differ in their translations**, and **what methods can effectively identify these differences?**

To answer these questions, we conduct a systematic comparison of LLM, NMT, and HT translations using explainability techniques (Lundberg and Lee, 2017; Rajagopal et al., 2021; Yin and Neubig, 2022; Wu et al., 2023), namely Leave-One-Out (LOO) (Li et al., 2016) and Integrated Gradients (IG) (Sundararajan et al., 2017). Specifically, we use a two-stage approach: first, we classify texts in the same target language based on their origin —whether they are original texts (O) written by native speakers or translations (T), whether human or automated. Next, we apply post-hoc interpretability methods to extract key features that contribute to these classifications. Our analysis focuses on identifying whether the most important features for O/T classification are consistent across LLM-based, NMT-based, and human translation outputs.

To understand these distinctions, we perform two analyses: (i) Feature Overlap Analysis: we calculate the average intersection of the top most important lexical features used across different translation systems to classify O/T, focusing on how much the most important features identified by explainability techniques overlap across LLM, NMT, and HT systems, and (ii) Feature Frequency Analysis: we analyse the frequency distribution of these key linguistic features within each translation system.

Our findings show that while many LLMs and NMT systems produce good translations, they still differ from content originally written by native speakers. LLMs like Aya-101-13B and TowerInstruct-7B-v0.2 exhibit alignment with traditional NMT models, such as DeepL and NLLB-600M, regarding O/T classification accuracy compared to content originally authored in the target language. Overall, our results confirm that NMT translations are more readily distinguishable from originals, with traditional NMT systems generally outperforming LLMs in translation quality and consistency. At the same time, human-generated translations remain distinctly different from those produced by machines.

Using explainability methods, we identified the key features that differentiate translations produced by LLMs, NMT systems, and human translators. Our findings suggest that LOO is generally better at pinpointing the most critical single feature, while IG is more effective when considering a broader range of important features. Moreover, our analysis shows that LLMs like Gemma-7B and TowerInstruct-7B-v0.2 often align closely with NMT systems such as M2M-100-418M and DeepL in their lexical feature selection during translation. Finally, our findings show that LLMs generally exhibit PoS patterns more aligned with HT than NMT models, particularly in the use of adverbs and auxiliary verbs. However, human translations consistently exhibit lower overlap with certain linguistic features from both LLMs and NMT systems, indicating that despite some shared patterns, human translations retain a unique quality.

The paper is structured as follows: Section 2 outlines our experimental design, and Sections 2.1 and 2.2 detail the data and models used in our study. Section 3 discusses our strategies for evaluation of translation quality and methods we employ for extracting important distinctive features of original and translated texts, while Section 4 examines the differences in classification features between LLMs, NMT systems, and human translations. Finally, Section 5 concludes the paper.

2 Experimental Design

To identify important explanations with respect to O/T classification in the outputs of translation systems, we apply explainability methods to each sentence and generate attribution scores for the tokens. Below, we describe the methods used to produce these attribution scores.

Leave-One-Out (LOO). We use LOO (Li et al., 2016), a popular model-agnostic feature attribution technique, to compute the attribution score for each token x_i in an input sentence X with respect to the model's prediction $\hat{y}$. Let $w_{[\text{CLS}]}$ be the final layer representation of the "[CLS]" token for X. During inference, the method processes the input through ReLU, affine, and softmax layers to produce a probability distribution over the outputs. For each token x_i, LOO measures the change in probability when x_i is excluded from the input X. Higher change in probability indicates that the token x_i is more influential in the model's prediction:

$$\ell = \text{softmax}(\text{affine}(\text{ReLU}(w_{[\text{CLS}]})))$$

$$\ell_i = \text{softmax}(\text{affine}(\text{ReLU}(w_i)))$$

$$\nabla_i = \ell - \ell_i$$

where w_i represents the final layer output of the "[CLS]" token when the token x_i is removed from the input sequence X.

Integrated Gradients (IG). Sundararajan et al. (2017) propose this technique for attributing a neural network's output to its input features by computing the integral of the gradients of the model's prediction with respect to the inputs along a path from a baseline to the actual input. The attribution for a feature x_i is given by:

$$\text{IG}_i = (x_i - x_i^0) \cdot \int_0^1 \frac{\partial f(x^0 + \alpha \cdot (x - x^0))}{\partial x_i} \, d\alpha$$

where x_i^0 is the baseline input and f is the model's prediction function.

In this work, IG is used to compute attribution scores for each token x_i[2] in X. IG provides scores between -1 and 1 for each embedding dimension of the token x_i, where 1 and -1 represent maximum influence towards labels 1 (T) and 0 (O), and scores near zero indicate minimal impact.

2.1 Data

We use the Monolingual German dataset from the Multilingual Parallel Direct Europarl (MPDE) featuring annotated paragraphs from the proceedings of the European Parliament (Amponsah-Kaakyire et al., 2021). The dataset includes both the original texts and their translations. Each paragraph, averaging 80 tokens, is labeled to indicate whether it is an original or a translation. Since most NMT systems operate on sentence level, we split each paragraph into sentences, which we later use for our work.

However, in MPDE, paragraphs of German sources typically contain more sentences than their English translations.[3] To address this imbalance, we remove certain amount of German source sentences, creating a training set with an equal number of original and translated sentences (97,108 in the training set and 20,744 in the test set).

To further perform evaluation of translation quality, we need a clear one-to-one correspondence between source sentence, human-translated sentence and the automatically translated sentence. As mentioned above, not every paragraph of the MPDE dataset has the same number of sentences in its German source and in its English translation. We have composed a subset of MPDE consisting only of those sentences whose paragraphs have an equal number of German and English sentences. This subset contains 38,035 sentences.

Pre-processing. To ensure that the explanation methods work efficiently, we tokenize and truecase our data.[4] Both are performed using Moses scripts (Koehn et al., 2007).

2.2 Models

We report O/T classification and translation quality results on a wide selection of some of the best-performing models, both commercial and open-source models:

- **DeepL Translator**: a state-of-the-art commercial NMT system.[5]
- **Google Translate**: Likely the most widely used commercial NMT system.[6]
- **M2M-100-418M** (Fan et al., 2020): A large multilingual NMT model trained on 2,200 translation directions, enabling many-to-many translation across 100 languages. We use the base version.
- **MADLAD-400** (Kudugunta et al., 2023): A multilingual NMT model based on the T5 architecture (Raffel et al., 2020b), with 3 billion parameters, trained on 1 trillion tokens across 450 languages using publicly available data.
- **NLLB-600M** (Costa-jussà et al., 2022): It represents the current state-of-the-art NMT system,

[2]Token x_i may refer to either a whole word or its subunits, as the WordPiece tokenizer (Song et al., 2021) splits words into subunits. To compute the attribution score at the word level, we average the attributions of its subunits.

[3]This is due to the fact that the translations of paragraphs are not aligned sentence-wise. While the original paragraph may have i sentences, one translation may have j sentences and another k.

[4]As further we need, for example, to analyze lexical overlaps, it is important that we do not miss out on words because of punctuation or case

[5]https://www.deepl.com/en/translator (accessed on August 16, 2024)

[6]https://translate.google.com/?sl=de&tl=en&op=translate (accessed on August 13, 2024)

System	O/T Classification Accuracy (%)	AEM	
		COMET	BLEU
HT	0.79		
DeepL	0.86	**0.85**	**34.85** ± 0.19
Google Translate	0.92	0.79	24.17 ± 0.16
M2M-100-418M	0.91	0.81	25.94 ± 0.16
MADLAD-400-MT	0.91	0.69	16.37 ± 0.18
NLLB-600M	0.83	0.79	27.35 ± 0.19
LLaMAX-3.1-8B-Alpaca	**0.94**	0.81	15.43 ± 0.13
TowerInstruct-7B-v0.2	0.83	**0.84**	33.35 ± 0.18
Aya-101-13B	0.86	0.83	25.35 ± 0.16
Gemma-7B	0.89	0.83	27.53± 0.19
Llama-3.1-IT-8B	0.90	0.82	26.91 ± 0.17

Table 1: Performance metrics for various systems including classification accuracy and automatic MT evaluation metrics (COMET and BLEU). The highest scores are highlighted in bold.

scaling up to 200 languages. We experiment with the distilled version with 600M parameters.

In addition to the NMT systems listed above, we pick three well-known and high-performing open-source LLMs and use them for prompt-based translation without any prior fine-tuning (see Appendix A for the prompt templates):

- **LLaMAX-3.1-8B-Alpaca** (Lu et al., 2024) is an open-source instruction-following language model with 8 billion parameters. It is fine-tuned from the LLaMA model (Taori et al., 2023) and supports 102 languages through continual pre-training, incorporating 52,000 Self-Instruct English instruction examples (Wang et al., 2023).
- **Llama-3.1-IT-8B** (Dubey et al., 2024): The Meta Llama 3.1 collection includes multilingual LLMs. This 8B parameter model is pretrained and instruction-tuned for text generation, optimized for multilingual dialogue.
- **TowerInstruct-7B-v0.2** (Alves et al., 2024): A language model based on LLaMA 2 (Touvron et al., 2023), using a diverse dataset of 20 billion tokens from monolingual sources in ten different languages.
- **Aya-101-13B** (Üstün et al., 2024): A 13-billion-parameter mT5 (Xue et al., 2021) multilingual model trained on instructions in 101 languages, exceeding the coverage of earlier open-source models (Lai et al., 2023; Muennighoff et al., 2022; Le Scao et al., 2023).
- **Gemma-7B** (Team et al., 2024b) is a lightweight open-source LLM developed by Google DeepMind. It has been instruction-tuned to respond to prompts in a conversational manner.

3 Evaluation

3.1 O/T Classification

We follow Dutta Chowdhury et al. (2022) to perform binary classification between original and translated (O and T) sentences. We use the XLM-RoBERTa base model (Conneau et al., 2020) with a softmax classifier applied to the `[CLS]` token of the sentence embeddings. We freeze hyperparameters and weights of the pre-trained encoder, and train the classifier for 10 epochs on each sentence with batch size of 16 and learning rate of 2×10^{-5}. All experiments are performed using NVIDIA V100 or A100 GPUs.

Results. The linear O/T classifiers show high accuracies (>80%) for all models (Table 1). We find that the automatically translated sentences, for both NMTs and LLMs, are always identified with higher accuracy than the human-translated ones. This finding corroborates the hypothesis that automatically translated texts are more readily distinguishable in classification tasks than those translated by humans (Ilisei et al., 2010; Rubino et al., 2016; Pylypenko et al., 2021).

3.2 Translation Quality

To assess translation quality, we utilise two automatic evaluation metrics (AEM): BLEU (Papineni et al., 2002) as implemented in SacreBLEU[7] (Post, 2018) and COMET (Rei et al., 2022).[8] BLEU relies on word n-gram similarity, whereas COMET

[7] BLEU signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0

[8] Unbabel/wmt22-comet-da, see `https://github.com/Unbabel/COMET`

System	LOO			IG		
	top-1	top-3	top-5	top-1	top-3	top-5
HT	0.64	0.66	0.66	0.51	0.56	0.57
DeepL	0.60	0.73	0.72	0.53	0.61	0.71
Google Translate	**0.78**	0.70	0.76	0.50	0.50	**0.83**
M2M-100-418M	0.57	0.70	0.76	0.57	0.75	0.75
NLLB-600M	0.50	0.73	0.69	0.58	0.50	0.71
TowerInstruct-7B-v0.2	0.54	0.70	0.74	0.51	0.55	0.54
Aya-101-13B	0.53	0.69	0.76	0.53	0.72	0.68
Gemma-7B	0.54	0.65	0.63	0.55	0.55	0.53
Llama-3.1-IT-8B	0.50	0.73	0.76	0.51	0.64	0.65
Mean	0.58	0.70	0.72	0.53	0.60	0.66

Table 2: Performance of the sufficiency classifier across different ranks (top-1, top-3, top-5) using LOO and IG methods for HT, NMT, and LLM systems. The highest scores for each method are highlighted in teal (LOO) and gray (IG), with the highest scores boldfaced to highlight the strengths of each method.

is a semantic metric built upon the XLM-R architecture.

Results. Table 1 shows that across different models, COMET scores remain relatively stable, while BLEU scores show greater fluctuation. DeepL stands out as the top performer, achieving the highest scores in both COMET (0.85) and BLEU (34.85). TowerInstruct-7B-v0.2 also performs well, particularly in COMET, reflecting high translation quality. Two systems, LLaMAX-3.1-8B-Alpaca and MADLAD-400-MT, exhibit poor translation quality. The high number of translation errors could skew the explainability results, focusing on these mistakes rather than models' intrinsic characteristics. Therefore, we exclude these models for further experiments. We perform a correlation analysis, and find no significant correlation between translation quality and O/T classification accuracy. See Appendix C for more details.

3.3 Do explanations capture sufficient information?

Understanding the effectiveness of model predictions often relies on the quality of explanations derived from those models. In this context, an explanation refers to the rationale behind a model's predictions, specifically identifying the *input tokens (features)* that most significantly influence the classification outcome. We follow the approach outlined by Xie et al. (2024) to evaluate the sufficiency of these explanations, as defined by Jacovi et al. (2018) and Yu et al. (2019). Sufficiency refers to the average change in predicted class probability when only the top k influential tokens are retained. This metric assesses how well the top k attributions explain the model's predictions, ultimately determining whether these explanations faithfully represent the model's decision-making process.

Previous research (Amponsah-Kaakyire et al., 2022) has shown that feature attribution including IG can be used to identify input tokens that are particularly important to O/T classification results for original texts and human translations.

However, whether this holds true across different types of translations, such as those generated by large language models (LLMs) or neural machine translation systems (NMT), remains underexplored. Bizzoni et al. (2020) investigated this problem using PoS perplexity scores and syntactic dependency lengths. More recently, Luo et al. (2024) systematically investigate the differences in the distribution of translation divergences between HT and MT through a large-scale, fine-grained comparative analysis, focusing on morphosyntactic variations. In contrast, our approach investigates lexical (words and PoS) differences by analysing explanations from O/T classifiers.

Our goal is to identify the key features that set apart translation artifacts produced by LLMs, NMT, and HTs from the text originally authored in the target language. To evaluate the sufficiency of our methods—specifically Leave-One-Out (LOO) and IG—we separately extract the top k tokens with the highest attribution scores for each sentence in the training set (see Section 2.1). We then construct datasets with sentences consisting only of these top k tokens while maintaining the same labels. O/T classifiers are then trained on these datasets,

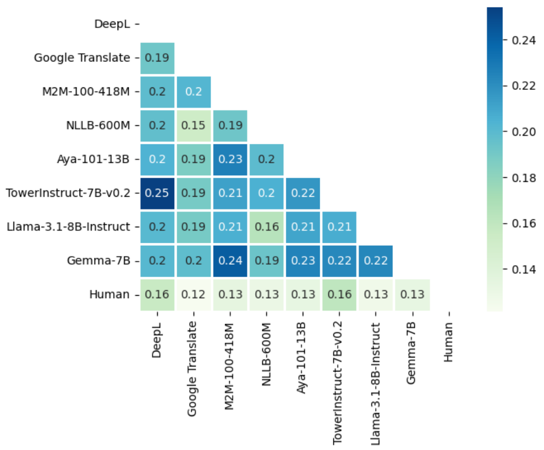

Figure 1: Level of intersection between top-5 most important explanations across different translation methods using LOO method.

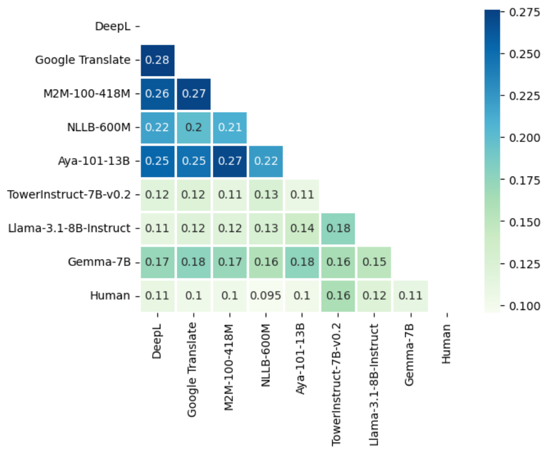

Figure 2: Level of intersection between top-5 most important explanations across different translation methods with IG.

where k = {1, 3, 5}, and we subsequently assess the classifiers' accuracy on the test set (Table 2) [9].

3.3.1 Sufficiency

If we can maintain high accuracy of O/T classifier using only the k tokens with the highest attribution scores, this indicates that the explainability methods (LOO and IG) work as intended, allowing us to efficiently identify important differences between translations and originally authored sentences in the target language.

Results. Table 2 shows that high accuracy for O/T is consistently maintained for the top k tokens with the highest attribution scores, indicating that the explainability methods (LOO and IG) function as intended. On average, as the number of tokens increases, we see an improvement in the sufficiency scores, indicating that the features we are extracting are indeed important.

Moreover, LOO is able to achieve much higher sufficiency score on top-1 tokens from certain model outputs as compared to IG, suggesting that LOO may be more effective at pinpointing the most critical token for classification. The reason for that might be that Leave-One-Out (LOO) directly removes each word and measures the impact on model prediction, giving a more precise attribution score. In contrast, Integrated Gradients (IG) require pooling attributions across the dimensions of an embedding and averaging attributions across subwords when a word is split into pieces, which may provide better performance in context, but lower it when focusing on a single word.

The LOO method achieves its highest top-1 sufficiency score of 0.78 across all models for Google Translate, underscoring its potential effectiveness in identifying essential tokens. In contrast, the IG method records its highest top-5 sufficiency score of 0.83 for the same translation system, showcasing its strength in capturing significant features across a broader range of tokens.

4 Feature Analysis of LLM, NMT, and Human Translation

4.1 Feature Overlap Analysis

We conduct an intersection analysis of linguistic features (input tokens), focusing specifically on sentences for which we can establish a one-to-one correspondence between outputs of different translation systems. For these sentences, we apply both LOO and IG using previously trained O/T classifiers for HT, NMT, and LLM datasets. This process enables us to compute attribution scores for individual tokens within each sentence. Using these scores, we extract the top-k most important tokens ($k = 1, 3, 5$) for each sentence.

Following this, we calculate the intersection between the LOO and IG results for different translation systems using the Jaccard Similarity Coefficient, which represents the percentage of common tokens and takes a value from 0 to 1. A high intersection among the top-k tokens indicates robust features (tokens) that are consistently identified as important across different translation models.

[9] We modified the train set for the sufficiency experiment but left the test set unchanged to ensure fair evaluation.

Conversely, if the intersection between systems and/or human translations is low, it indicates that the translations exhibit different features. Figure 1 presents the pairwise Jaccard values for the top-5 features derived from the Leave-One-Out (LOO) method. Each cell quantifies the degree of overlap between the top features of two different translation systems, with darker shades representing higher overlaps. Notably, the highest intersection is observed between TowerInstruct-7B-v0.2 and DeepL, with an overlap of 0.25, suggesting a strong similarity in the features identified for these models.

Another substantial intersection occurs between Gemma-7B and M2M-100-418M at 0.24, indicating considerable alignment in their outputs. In contrast, human-generated content shows relatively lower intersections with machine models, such as 0.16 with TowerInstruct-7B-v0.2 and DeepL and 0.13 with M2M-100-418M, underscoring the unique nature of human translations compared to machine-generated translations.

Similarly, Figure 2 shows the pairwise Jaccard values for the top-5 features (tokens) obtained using Integrated Gradients (IG). The most notable overlap is between Google Translate and DeepL, with a significant intersection of 0.28, demonstrating a strong similarity in their feature selections. A notable intersection of 0.27 is observed between M2M-100-418M and both Aya-101-13B and Google Translate, suggesting that these models yield quite similar results. The lower intersection of 0.11 between TowerInstruct-7B-v0.2 and Aya-101-13B emphasizes the differences in their outputs. The intersection with human translation identified by IG is notably highest for TowerInstruct-7B-v0.2, at a value of 0.16.

The combined results suggest that while certain LLMs, like Aya-101-13B and TowerInstruct-7B-v0.2, closely align with NMT models such as M2M-100-418M and DeepL in their feature selection, others retain unique classification features. Furthermore, there are notable differences in how closely these models align with human translations, with TowerInstruct-7B-v0.2 demonstrating the highest similarity to HT as shown by both LOO and IG.

4.2 Feature Frequency Analysis

We examine the frequency of different Part of Speech (PoS) tags across translation systems, focusing on the top k features flagged by LOO/IG for each sentence. For each system, we group sentences – both human and machine translations – into predefined sentence length bins. These bins are divided into ranges (e.g., 0-10, 10-15, 15-20 words), and for each, we calculate and normalize the frequency of the identified features based on the total number of sentences in that bin. This helps us compare trends in PoS distribution as sentence length increases. We are examining trends for the 9 most common PoS.

To ensure the reliability of our measurements, we account for the margin of error (standard deviation) obtained through bootstrapping by subsampling each bin 1,000 times while maintaining the PoS distribution within each sentence. In the graphs we show the standard deviation with shading. Figure 3 illustrates variations in PoS distribution, showing nine subplots for adverbs (ADV), verbs (VERB), determiners (DET), auxiliary verbs (AUX), nouns (NOUN), pronouns (PRON), adjectives (ADJ), adpositions (ADP), and punctuations (PUNCT).

For ADV, most models—both NMT and LLM—use fewer adverbs than HT. However, Llama-3.1-8B demonstrates frequencies that are closer to HT as sentence length increases, while TowerInstruct-7B-v0.2 diverges with longer sentences. NMT models like M2M and Google Translate underproduce ADV compared to HT, whereas DeepL aligns more closely with HT and tends to overproduce ADV with longer sentences.

ADP use in HT increases with sentence length, and most NMT and LLM models follow this trend, although models like Google Translate show slightly lower frequencies in longer sentences. Pylypenko et al. (2021) find that the relative frequencies of ADV and ADP in PoS tagging are strong indicators of translationese in HT.

For VERB, both HT and most NMT and LLM models maintain a steady frequency, though the models generally underproduce compared to the human translation trend. For DET, HT usage slightly increases with sentence length, while all LLM and NMT models, except DeepL, tend to use determiners more frequently.

In the case of PRON, most models tend to align with the human trend for shorter sentences. However, as sentence length increases, their frequencies start to deviate from each other. NLLB-600M demonstrates a substantially higher frequency than human translations across all sentence lengths.

In ADJ usage, HT remains relatively stable,

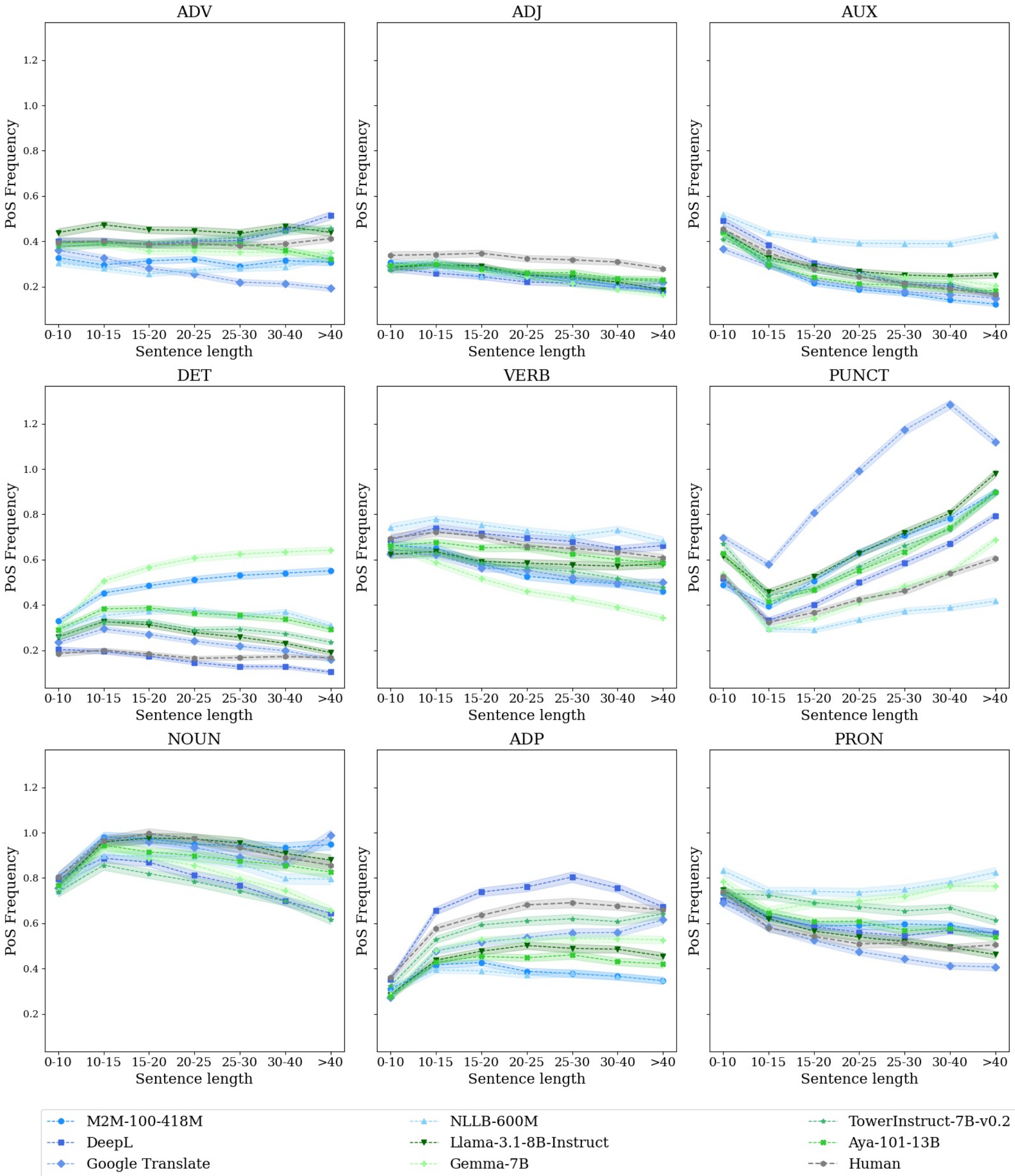

Figure 3: The frequency of the top PoS categories flagged by LOO across different sentence length bins. The x-axis of each subplot represents sentence length, divided into ranges (0-10, 10-15, 15-20, etc.), and the y-axis shows PoS frequency, indicating how often each PoS occurs in sentences of different lengths.

showing a slight decrease as sentence length increases. All NMT and LLM models exhibit lower adjective frequencies overall, with their trends being extremely similar across all sentence lengths.

For AUX, HT demonstrates a consistent decline as sentence length increases. Most NMT models follow this trend, except for NLLB-600M, which shows significantly higher AUX usage. Similarly, Llama-3.1-8B-Instruct exhibits slightly higher AUX frequencies compared to HT. The frequency of NOUN usage is maximal for sentences of length 10-15 and then consistently decreases for longer sentences. HT and most models seem to follow this trend, except for two NMTs (M2M-100 and Google Translate), which tend to overproduce nouns in very long sentences. For HT and NMT/LLM, the frequency of PUNCT usage in sentences of length 10-15 is lower than in shorter sentences, although there is an increasing trend for sentences longer than 15. Google Translate exhibits notably higher PUNCT frequencies than all other models and HT, although its usage declines in very long sentences.

Overall, LLMs exhibit PoS patterns (for 6 out of 9 tags) that closely align with human translations, whereas NMT models show greater deviations, particularly regarding PUNCT. NMT models tend to underproduce ADV, and for some other parts of speech (PoS) like ADP or PRON, they show significant divergence. In contrast, LLMs exhibit stronger agreement in trends and align more closely with HT, although they still demonstrate some overuse in short sentences. Both NMTs and LLMs underproduce ADJ compared to HT, particularly in longer sentences. LLMs better mimic human usage in ADV and AUX frequencies, especially in longer sentences. Appendix B displays the frequency plots of the top PoS categories identified by Integrated Gradients (IG) across various sentence-length bins.

5 Conclusion

In this work, we systematically explore the translation divergences between LLMs, NMTs, and human translations. Our key findings show distinct differences in how these systems approach translation, despite advancements in LLMs that allow them to produce high-quality outputs. We find that while LLMs often exhibit translation patterns more similar to human translations compared to traditional NMT models, they still diverge from originally authored text in the same language. Overall, we find that automatically translated sentences from both NMTs and LLMs are consistently identified with higher accuracy in O/T classification tasks than human-translated ones. This supports the hypothesis that machine-translated texts are more easily distinguishable from original texts than those translated by humans (Rubino et al., 2016; Pylypenko et al., 2021).

To better understand the distinctions between translations produced by LLMs and NMTs compared to human translations, we employ Leave-One-Out and Integrated Gradients explanation methods to extract and analyze lexical features identified by translation classifiers. Our findings indicate that even when using a sufficiency-based approach, we can recover a significant amount of O/T classification accuracy. This demonstrates that these features are effective in distinguishing between automatic and human translations.

Further, our results indicate that sufficiency-based approach is particularly effective at identifying single critical features, while Integrated Gradients (IG) capture a broader range of important features. Interestingly, we observe that certain LLMs align closely with NMT systems in their feature selection, demonstrating similarities in their approaches. However, human translations consistently exhibit lower overlap with both LLM and NMT outputs, particularly regarding crucial features like punctuation and specific PoS.

Furthermore, our frequency analysis of PoS tags reveals that LLMs align more closely with HT in their usage, especially in terms of adverbs, and auxiliary verbs, while NMT models tend to overproduce specific tags in shorter sentences. This suggests that LLMs, although not perfect, are making strides in mimicking human translation patterns. Our findings highlight the characteristics that define the outputs of various translation systems. However, despite advances in machine translation, human translations continue to display distinctive characteristics, particularly in their nuanced use of linguistic features, making them less prone to the artifacts seen in machine-generated texts.

Limitations

Limitations of Lexical Features. The results presented in this study rely entirely on the lexical features derived from Leave-One-Out (LOO) and Integrated Gradients (IG), which may fall short of capturing the intricacies of translation quality. Moreover, translation artifacts can arise at both syntactic and semantic levels (Bizzoni et al., 2020; Briakou and Carpuat, 2020), aspects that this research does not address. This leaves an exploration of these dimensions to future work.

Prompting Choice. Prompting has demonstrated varying sensitivity to the choice of templates and examples (Zhao et al., 2021). In machine translation (MT), prior studies have used different templates (Brown et al., 2020; Chowdhery et al., 2023; Wei et al., 2021). In our work, we reevaluate these templates to determine the optimal one. However, the format and wording of the prompt significantly influence how the LLM comprehends the task and performs translation, potentially impacting our findings, which we leave for future exploration.

Stability of Model Outputs. Additionally, we have assumed that the output of a specific model would remain stable throughout the analysis. However, LLMs are frequently updated, which can lead to changes in their writing style and coherence. Such variations might cause explainability methods to underperform, exacerbating the issues discussed in this work.

Constraints of Sentence-Level Analysis. Most NMT models utilized in this study function effectively at the sentence level, necessitating that we translate individual sentences for both NMTs and LLMs to ensure consistency. Thus, our sentence-based analysis with LLMs is also a limiting factor, as it restricts our ability to capture broader contextual nuances (Koneru et al., 2024). This would entail expanding our analysis beyond sentence-level assessments.

Acknowledgments

We would like to thank Vagrant Gautam for their helpful feedback. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1102 Information Density and Linguistic Encoding.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Cristina España-Bonet, and Josef van Genabith. 2021. Do not rely on relay translations: Multilingual parallel direct Europarl. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 1–7, online. Association for Computational Linguistics.

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Josef Genabith, and Cristina España-Bonet. 2022. Explaining translationese: why are neural classifiers better and what do they learn? In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 281–296, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Eleftheria Briakou and Marine Carpuat. 2020. Detecting fine-grained cross-lingual semantic divergences without supervision by learning to rank. *arXiv preprint arXiv:2010.03662*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Dun Deng and Nianwen Xue. 2017. Translation divergences in Chinese–English machine translation: An empirical investigation. *Computational Linguistics*, 43(3):521–565.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. *arXiv preprint arXiv:2004.06063*.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In *Computational Linguistics and Intelligent Text Processing: 11th International Conference, CICLing 2010, Iaşi, Romania, March 21-27, 2010. Proceedings 11*, pages 503–511. Springer.

Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018. Understanding convolutional neural networks for text classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–65, Brussels, Belgium. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual refinement of translations: Large language models for sentence and document-level post-editing.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages. *arXiv preprint arXiv:2407.05975*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Jiaming Luo, Colin Cherry, and George Foster. 2024. To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation. *Transactions of the Association for Computational Linguistics*, 12:355–371.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Dmitry Nikolaev, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Maria Saeboe, and Omri Abend. 2020. Fine-grained analysis of cross-linguistic syntactic divergences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1159–1176, Online. Association for Computational Linguistics.

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

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

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H Hovy, and Yulia Tsvetkov. 2021. SELFEXPLAIN: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 836–850, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan Awadalla. 2023. Do gpts produce less literal translations? *arXiv preprint arXiv:2305.16806*.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef Van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 960–970.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast WordPiece tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh

Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024b. Gemma 2: Improving open language models at a practical size.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. *arXiv preprint arXiv:2210.14250*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. *arXiv preprint arXiv:1803.11112*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. 2023. Interpretability at scale: Identifying causal mechanisms in alpaca. In *Advances in Neural Information Processing Systems*, volume 36, pages 78205–78226. Curran Associates, Inc.

Roy Xie, Orevaoghene Ahia, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Extracting lexical features from dialects via interpretable dialect classifiers. *arXiv preprint arXiv:2402.17914*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. *arXiv preprint arXiv:2202.10419*.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

A Prompts

LLaMAX-3.1-8B-Alpaca

```
Below is an instruction that describes a task, paired with an input that provides
further context.
Write a response that appropriately completes the request.
### Instruction: Translate the following sentences from {source} to {target}.
Input:
{input_sentence}
### Response:
```

TowerInstruct-7B-v0.2

```
Translate the following sentence into {target}.
{source}: {input_sentence}
{target}:
```

Aya-101-13B

```
Translate to {target}: {input_sentence}
```

LLaMA-3.1-IT-8B

```
Translate the following sentence from {source} to {target}:
{input_sentence}
{target}:
```

Gemma-7B

```
Translate this sentence from {source} to {target} without any comments:
{source}:
{input_sentence}
{target}:
```

B

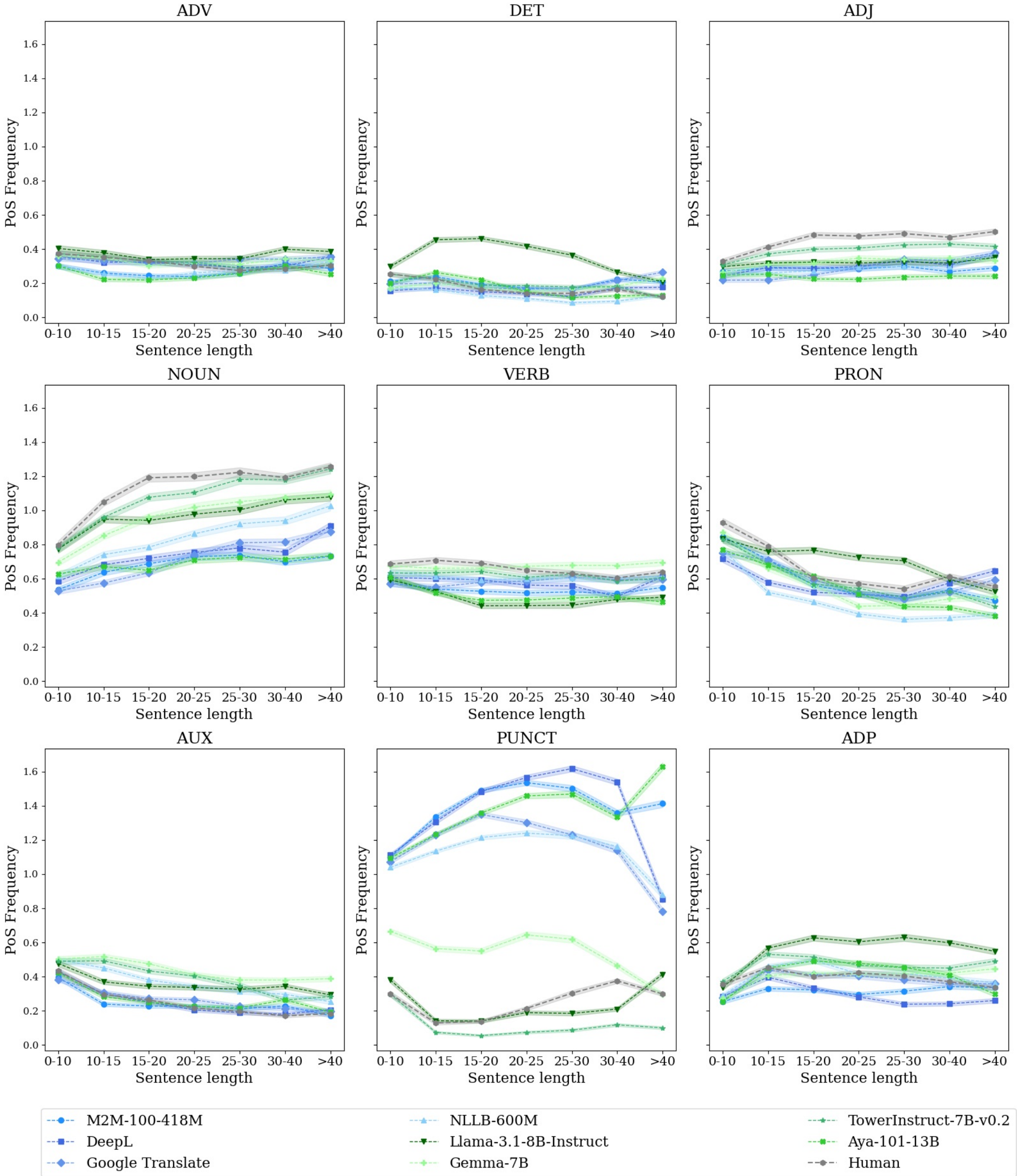

Figure 4: The frequency of the top PoS categories flagged by IG across different sentence length bins. The x-axis of each subplot represents sentence length, divided into ranges (0-10, 10-15, 15-20, etc.), and the y-axis shows PoS frequency, indicating how often each PoS occurs in sentences of different lengths.

C Correlation Analysis

We calculate Spearman's correlation to analyze the relationship between translation quality and O/T classification accuracy, considering a significance level $\alpha = 0.05$. We find Spearman's correlation between COMET and Accuracy to be -0.43 with p-value 0.28, and -0.63 with p-value 0.1 between BLEU and Accuracy. Correlations are not statistically significant; therefore, given our data, there is no evidence to support the notion that poorer translations are more easily classified as translated or non-translated texts.