# From Shortcuts to Balance: Attribution Analysis of Speech-Text Feature Utilization in Distinguishing Original from Machine-Translated Texts

**Yongjian Chen**[*]
Center for Language and Cognition
University of Groningen
Groningen, Netherlands
yongjian.chenl@rug.nl

**Antonio Toral**
Dep. de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
Sant Vicent del Raspeig, Spain
antonio.toralr@ua.es

## Abstract

Neural text-based models for detecting machine-translated texts can rely on named entities (NEs) as classification shortcuts. While masking NEs encourages learning genuine translationese signals, it degrades the classification performance. Incorporating speech features compensates for this loss, but their interaction with NE reliance requires careful investigation. Through systematic attribution analysis across modalities, we find that bimodal integration leads to more balanced feature utilization, reducing the reliance on NEs in text while moderating overemphasis attribution patterns in speech features.

## 1 Introduction

The development of neural machine translation (NMT) architectures has notably improved translation accuracy (Chen et al., 2018; Team et al., 2022). Despite these gains, NMT models still exhibit distinguishable characteristics from human translation (Toral et al., 2018). NMT outputs are characterized by machine translationese—a distinctive linguistic phenomenon that diverges from real target language use (Vanmassenhove et al., 2021; Dutta Chowdhury et al., 2022). This artificially constructed language manifests at various linguistic levels, affecting lexical choice, syntactic structure, semantic interpretation, discourse coherence, etc. in the target language domain (Vanmassenhove et al., 2021). These observable differences have enabled researchers, e.g. Jones and Sheridan (2015); Iyyer et al. (2018), to develop automatic methods for distinguishing machine-translated texts from those originally authored in the target language. Such detection is increasingly important as MT systems become widely used: for content authenticity verification in news and documents, quality assurance in professional translation workflows, research integrity in linguistic studies, and preventing malicious use of MT for generating inauthentic content.

Recent work has demonstrated strong performance in distinguishing machine-translated text from human-authored or human-translated content using pretrained language models (Pylypenko et al., 2021; van der Werff et al., 2022). However, attribution analysis through Integrated Gradient (IG) (Sundararajan et al., 2017) has exposed a critical limitation in these models: they rely on superficial patterns, such as named entities (NEs), rather than on genuine translationese signals alone (Amponsah-Kaakyire et al., 2022). This behavior, termed "Clever Hans" by Borah et al. (2023), suggests that these models exploit spurious correlations instead of capturing deeper linguistic cues. NE reliance is problematic, as it represents topic-based shortcuts rather than genuine translationese features—robust models should detect translationese regardless of content domain. As an example, if an English text is translated from German, we argue that it would be more principled to classify it as a translation due to its linguistic features (e.g. its syntax is somewhat German-like), rather than the fact that it contains NEs that refer to locations in a German-speaking country (e.g. Berlin).

To mitigate this issue, Borah et al. (2023) masked NEs during training. While this approach reduces reliance on NEs, it also results in a 2.6-3.2 percentage point drop in classification accuracy. More recently, Chen et al. (2025) have explored integrating speech features alongside text-based models to compensate for this performance decline. Both handcrafted and self-supervised speech representations have shown promise in preserving accuracy despite NE masking the text modality.

Amponsah-Kaakyire et al. (2022) and Borah et al. (2023) primarily analyzed attribution at the subword token level and relied on averaged fea-

[*] Work partially done during a research visit at Universitat d'Alacant

*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32756–32763
November 4-9, 2025 ©2025 Association for Computational Linguistics

ture attributions, potentially overlooking important linguistic patterns at the word or entity level. Moreover, attribution analyses in Borah et al. (2023) largely depend on ranked token lists, providing an incomplete perspective on model behavior. Additionally, while the bimodal classifiers in Chen et al. (2025) show promise for maintaining performance despite NE masking, it remains unclear whether they truly address the underlying attribution bias through cross-modal dependencies. Understanding how speech and text features interact in their reliance on spurious correlations is crucial for developing genuinely robust detection systems that avoid topic-based shortcuts regardless of modality combination.

To address these gaps, we propose a comprehensive attribution analysis framework that systematically investigates model behavior across modalities, particularly aiming to better understand how bimodal classifiers balance feature utilization for machine translationese classification while mitigating spurious correlations. Our work[1] makes two main contributions: (1) introducing a systematic attribution analysis framework that enables word and entity-level comparisons across modalities, (2) quantifying how bimodal integration balances feature utilization by showing reduced NE reliance in text while moderating overly strong attribution patterns in speech features.

## 2 Proposed Framework

We employ IG to analyze model behavior in translationese classification across text, speech, and bimodal settings. Our analysis framework consists of attribution computation and multiple aggregation strategies at different granularity levels.

### 2.1 Attribution Computation

**Unimodal Setting** For the text modality, we compute attributions with respect to input embeddings, yielding an attribution tensor $A \in \mathbb{R}^{m \times n}$ where $m$ stands for sequence length and $n$ for feature dimension. For the speech modality, the attributions are computed on the raw audio features, producing a tensor $B \in \mathbb{R}^{s}$ where $s$ is the sequence length. We use Captum's IG implementation (Kokhlikyan et al., 2020) with 50 steps for integral approximation, using zero tensors as baselines for text and silence tensors for speech.

**Bimodal Setting** For bimodal inputs, we compute attributions for each modality while keeping the other fixed. Given a text-speech pair processed through their respective encoders, we calculate: (i) Text attributions with respect to input embeddings ($T \in \mathbb{R}^{m \times n}$) while keeping audio features fixed; (ii) speech attributions with respect to raw audio features ($S \in \mathbb{R}^{s}$) while keeping text embeddings fixed. The computation yields separate attribution tensors for each modality: a text attribution tensor $T \in \mathbb{R}^{m \times n}$ and a speech attribution tensor $S \in \mathbb{R}^{s}$, maintaining the same dimensionalities as in the unimodal setting. This way we can analyze how each modality's contribution is influenced by the presence of the other one. This allows us to subsequently analyze each modality's contribution to the model's decisions in the same way as the unimodal case, while accounting for the cross-modal effects on attribution patterns.

### 2.2 Attribution Aggregation

To investigate potential spurious correlations in translationese classification, particularly how models may rely on topical information from NEs versus non-NEs, we need to compute feature attribution scores at two levels: for individual word spans (aggregating attributions for each word) and for entity spans (aggregating attributions for each NE or non-NE sequence). For example, in *"The Berlin Wall fell in November 1989,"* we first aggregate subword tokens, e.g. ([Ber] [##lin]) into words ([Berlin]), and then group words into individual NE spans ([Berlin Wall], [November 1989]) versus individual non-NE spans ([The], [fell], [in]).

The aggregation process differs between the text and speech modalities. For the text modality, we first aggregate attributions across tokens within each span (word or entity) for each feature dimension using one of two approaches: (i) a mean-based method that captures overall feature contributions by averaging across tokens, or (ii) a signed-max method that identifies influential features by selecting the maximum magnitude value while preserving its sign. In contrast, speech modality attribution tensors do not require this step as there is no feature dimension to consider (as described in Section 2.1).

Subsequently, for both text and speech modalities, we apply one of three normalization variants to obtain scalar scores: (i) mean normalization, which captures the central tendency of attributions; (ii) L1 normalization, which sums absolute values to track total magnitude regardless of sign; and (iii)

[1] Code and data available at: https://github.com/yongjianchen-lang/bimodal_mt_discrimination

| Split | English |  | Chinese |  |
|---|---|---|---|---|
|  | Orig. | MT | Orig. | MT |
| **Training** |  |  |  |  |
| Total | 5,938 | 5,237 | 5,481 | 5,237 |
| w/ NE | 4,243 | 3,630 | 4,205 | 3,394 |
| w/o NE | 1,695 | 1,607 | 1,276 | 1,843 |
| **Test** |  |  |  |  |
| Total | 1,500 | 1,498 | 1,498 | 1,498 |
| w/ NE | 1,046 | 985 | 1,215 | 910 |
| w/o NE | 454 | 513 | 283 | 588 |

Table 1: Dataset distribution for English and Chinese splits showing original and machine-translated sentence counts.

L2, which applies the Euclidean norm to give more weight to features with larger attribution values.

## 3 Experimental Setup

**Data** Following Chen et al. (2025), we use WMT news task data[2] for English and Chinese targets, translating German source texts from WMT's German-to-English datasets into both target languages via Google Translate. This setup compares English (Germanic, non-tonal) with Chinese (non-Germanic, tonal). Training data spans 2014–2017 (English) and 2017–2019 (Chinese), with 2018 and 2020 as test sets, respectively. The final dataset composition is shown in Table 1. Speech data is synthesized by employing Microsoft Azure TTS API's male voices (en-US-AndrewNeural for English, zh-CN-YunyangNeural for Chinese). Model accuracy is evaluated on the complete test set, and attribution analysis concerns solely sentences containing NEs.

**Text Classifiers** We use DeBERTA-v3-large (He et al., 2021) and MacBERT-large (Cui et al., 2020) for English and Chinese respectively, following Chen et al. (2025). We fine-tune them on our training sets to detect machine-translated content under two conditions: standard fine-tuning and fine-tuning with NE masking. We use NER models *en_core_web_sm* (for English) and *zh_core_web_sm* (Chinese) from spaCy (Honnibal et al., 2020), replacing identified NEs in the training sets with the [MASK] token. The standard fine-tuning and the NE-masked fine-tuning altogether yield four text-based classifiers: *DeBERTa_ft* and *DeBERTa_ft_mask* for English, and *MacBERT_ft* and *MacBERT_ft_mask* for Chinese.

**Speech Classifiers** We maintain the use of *hubert-large-ll60k* (Hsu et al., 2021) for English and *chinese-hubert-large* (Guo and Liu, 2022) for Chinese from Chen et al. (2025). These models share architectural similarities with the text-based models, with the key distinction that HuBERT incorporates feature extraction layers preceding the transformer layers. Rather than following the previous two-stage approach (Chen et al., 2025), we adopt a streamlined approach, facilitating direct computation of integrated gradients attribution and providing better interpretability of model predictions by tracking feature importance back to the input. The models are directly fine-tuned under two parameter configurations: with frozen base parameters to mimic the two-stage implementation from Chen et al. (2025) and with unfrozen parameters to probe whether standard fine-tuning can lead the speech-based classifiers to rely more on NEs, as observed for text-based classifiers. Notably, in both parameter settings, the feature extraction layers' parameters remain frozen. Altogether we obtain four speech-based classifiers: *HuBERT_froz_en* and *HuBERT_unfroz_en* for English and *HuBERT_froz_zh* and *HuBERT_unfroz_zh* for Chinese.

**Bimodal Classifiers** Chen et al. (2025) demonstrated that integrating representations from fine-tuned BERT models (DeBERTa and MacBERT) with pre-trained HuBERT representations effectively compensates for the performance degradation caused by NE masking. Based on their probing results of the speech modality, we proceed to fuse DeBERTa and MacBERT, that are fine-tuned with NE-masking, with frozen HuBERT. In contrast to their cascade approach for training bimodal classifiers, we use the streamlined approach, so that we can compute IG attribution and maintain comparability across modalities. Given the robust performance for both target languages (Chen et al., 2025), we adopt the MISA fusion technique (Hazarika et al., 2020), which projects each modality into modality-invariant (capturing cross-modal commonalities through distributional alignment) and modality-specific (preserving unique modality characteristics) subspaces. This factorized approach is particularly well-suited for investigating attribution balance, as it enables us to separately analyze how models utilize shared cross-modal patterns versus modality-unique features, providing clearer insights into how bimodal integration affects reliance on spurious correlations. We train

[2] E.g. statmt.org/wmt20/translation-task.html

| Model | EN | ZH |
|---|---|---|
| DeBERTa_ft | 86.39 | - |
| DeBERTa_ft_mask | 84.32 | - |
| MacBERT_ft | - | 92.62 |
| MacBERT_ft_mask | - | 92.02 |
| HuBERT_froz_* | 60.64 | 71.50 |
| HuBERT_unfroz_* | 69.98 | 83.95 |
| MISA_mask_froz_* | 85.06 | 92.19 |

Table 2: Classification accuracy for English (EN) and Chinese (ZH). '*' represents the respective language.

two bimodal classifiers: *MISA_mask_froz_en* and *MISA_mask_froz_zh*.

**Attribution Comparison Configuration** We evaluate the classification accuracy on the complete test sets and analyze their behavior in distinguishing original from machine-translated texts via our proposed framework on sentences containing NEs, which is performed separately for each aggregation method and granularity level (see Section 2.2). For sentences, word and entity boundaries are both identified using the aformentioned SpaCy models, while for utterances, frames are aligned to word or entity spans using time stamps.[3]

To quantify the relationship between NEs and model attribution patterns, we compare the mean attribution scores ($\mu$) of NE spans versus non-NE spans across different models. We calculate a normalized difference score ($\delta$) between these means for each model, then measure the change in this difference ($\Delta$) between experimental conditions (standard vs. masked text, unfrozen vs. frozen speech parameters, and unimodal vs. bimodal inputs).[4] This analysis reveals how different training approaches affect the model's reliance on NEs.

## 4 Results

### 4.1 Classification Performance

Table 2 presents accuracy results across text-based, speech-based, and bimodal machine translationese classification models. The textual classifiers maintain performance levels consistent with those in the previous work, as expected given identical experimental configurations. While we observe some performance degradation in speech and bimodal classifiers under the streamlined approach compared to the two-stage implementation in Chen et al. (2025), this decline does not compromise our primary objective of analyzing attribution patterns. Importantly, the relative performance trends across experimental configurations align with their findings, validating our proposed analytical framework.

The speech-based models exhibit strong discriminative capacity, achieving classification accuracy substantially above chance level for both target languages. This confirms the effectiveness of acoustic features in distinguishing between original and machine-translated texts. The unfrozen fine-tuning configuration yields superior performance compared to its frozen counterpart. Moreover, the bimodal classifiers consistently outperform their NE-masked text-based counterparts across both languages, with particularly pronounced gains in English—replicating the previous findings in Chen et al. (2025). This consistent cross-lingual performance differential, despite the overall accuracy reduction under the streamlined approach, reinforces the robustness of the observed phenomena and underscores the compensatory effect of speech representations for NE masking in machine translationese classification.

### 4.2 Attribution Importance

Table 3 shows the results of the attribution analysis. For each condition, level and language we report the result with the approach that leads to the highest absolute score, which most often is signed-max aggregation and mean normalisation.[5]

**Attribution Analysis for Text Classifiers** Our analysis reveals consistently higher attribution contrasts between NE and non-NE spans in standard conditions compared to masked conditions for both languages, demonstrating that NE masking effectively reduces the model's reliance on NEs. This effect is substantially more pronounced in English ($\Delta$ 1.32) than in Chinese (0.25). Moreover, entity-level analysis consistently produces stronger standard-masked differences than word-level analysis across both languages, indicating that the text-based models more effectively recognize and utilize complete NEs compared to their component words.

**Attribution Analysis for Speech Classifiers** The comparative differences in Table 3 also capture a substantial shift in attribution patterns from unfrozen to frozen speech models. The positive $\Delta$ values across both languages at both levels confirm that unfrozen speech classifiers attribute relatively more importance to NEs than frozen classi-

[3] Obtained with github.com/readbeyond/aeneas

[4] Methodological details are provided in Appendix A.1.

[5] The complete results are provided in Appendix A.2.

| Modality | Experimental Condition (1st → 2nd) | Word Level |  | Entity Level |  |
|---|---|---|---|---|---|
|  |  | ZH | EN | ZH | EN |
| Text | Standard → Masked | 0.16 | 0.75 | 0.25 | 1.32 |
| Speech | Unfrozen → Frozen | 3.95 | 1.37 | 3.77 | 1.54 |
| Text | Unimodal Masked → Bimodal Masked | 0.08 | 0.21 | 0.11 | 0.28 |
| Speech | Unimodal Frozen → Bimodal Frozen | -0.06 | -1.26 | -0.09 | -1.21 |

Table 3: Key Attribution Pattern changes ($\Delta$) across experimental conditions. Positive values indicate decreased relative attribution to NEs in the second condition compared to the first; negative values indicate increased relative attribution to NEs in the second condition compared to the first.

fiers do. Therefore, although the unfrozen classifiers achieve better classification performance than their frozen counterparts (see Table 2), their highly stronger dependence on NEs raises concerns. This motivates our decision to utilize the frozen models for bimodal fusion. Unlike the text modality, speech-based models show comparable effects at both word and entity levels, suggesting a more uniform processing of linguistic units across different granularities than their text-based counterparts.

**Attribution Analysis for Bimodal Classifiers** The third result row of Table 3 shows the comparative differences between unimodal and bimodal settings for NE-masked text features. The unimodal text modality exhibits higher attribution contrasts than its bimodal counterpart for both languages, with stronger effects in English, suggesting that the unimodal-bimodal transition further reduces reliance on NE cues for the text modality. Additionally, both languages show greater unimodal-bimodal differences at the entity level than at the word level, indicating that text modality in bimodal classifiers, like their unimodal counterparts, processes complete NEs more effectively than their component words.

For speech features, the last row of results from Table 3 shows the comparative differences between unimodal and bimodal settings under frozen conditions. English displays negative $\Delta$, indicating that the bimodal integration increases the speech modality's relative attribution to NEs compared to the unimdoal frozen model, but still maintains a shift in smaller magnitude compared to the unimodal unfrozen-frozen contrast (compare $|\Delta|$ in the second result row of Table 3 to $|\Delta|$ in the last result row of Table 3). In contrast, Chinese displays minimal changes (small negative values of -0.06 at word level and -0.09 at entity level), suggesting that speech features maintain similar attribution patterns in both unimodal and bimodal settings.

These contrasting behaviors reveal complementary cross-modal patterns. In English, text features' reduced NE reliance is balanced by speech features shifting away from non-NE focus, creating more distributed attribution. In Chinese, where text changes are minimal, speech features maintain their non-NE emphasis. This suggests that the magnitude of change in text modality influences the balancing behavior in speech features, resulting in different strategies for bimodal feature utilization. The bimodal model can potentially learn to adaptively balance feature attribution between modalities, moderating overemphasis on specific feature types.

## 5 Conclusion

We examine how machine translationese classifiers process NEs across text, speech, and bimodal approaches, revealing distinct and robust patterns that remain consistent across different attribution aggregation methods. Text-based models show stronger entity-level than word-level attribution, with NE masking effects more pronounced in English than Chinese. Speech models exhibit contrasting behavior between parameter settings: unfrozen models emphasize NEs while frozen models favor non-NEs. Crucially, our bimodal analysis reveals that multimodal integration leads to more balanced attribution patterns - moderating text's reliance on NEs while also balancing the tendency of frozen speech features to overemphasize either NE or non-NE information. These findings suggest that multimodal approaches not only improve classification performance but also lead to more robust feature utilization.

## Limitations

Our work presents several opportunities for future research. First, while our analysis provides valuable insights using one widely-used translation system, investigating additional NMT architectures could reveal interesting variations in feature utilization patterns across different systems. Second, our speech synthesis methodology employs specific TTS voices (male voices for both languages); exploring a spectrum of voices with varied accents, genders, and emotions might offer additional perspectives on how speech features interact with named entities in translationese detection. Alternative data sources like audio books and dubbed movies with their accompanying transcripts could provide naturalistic speech-text pairs for extending this analysis. Third, expanding beyond German-to-English and German-to-Chinese translation directions could enhance our understanding of how attribution patterns manifest across more diverse language pairs and typological relationships. These extensions would complement our current findings and potentially reveal additional nuances in multimodal translationese classification.

## Acknowledgments

We gratefully acknowledge the financial support provided by the China Scholarship Council (CSC), which made this research possible. We extend our sincere thanks to our colleague Gabriele Sarti for his valuable advice on this work. We are also deeply appreciative of the anonymous (meta)reviewers, whose thoughtful comments and suggestions improved both the quality and clarity of this paper.

## References

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.

Angana Borah, Daria Pylypenko, Cristina España-Bonet, and Josef van Genabith. 2023. Measuring spurious correlation in classification: “clever hans” in translationese. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 196–206, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86, Melbourne, Australia. Association for Computational Linguistics.

Yongjian Chen, Mireia Farrús, and Antonio Toral. 2025. The potential of speech features to discriminate between original and machine-translated texts. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668.

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*, Seattle, United States. Association for Computational Linguistics.

Pengcheng Guo and Shixing Liu. 2022. chinese_speech_pretrain. Available online at: `https://github.com/TencentGameMate/chinese_speech_pretrain`, last accessed on 20-05-2025.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM ’20, page 1122–1131, New York, NY, USA. Association for Computing Machinery.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Jones and Lynnaire Sheridan. 2015. Back translation: an emerging sophisticated cyber strategy to subvert advances in ‘digital age’ plagiarism detection and prevention. *Assessment & Evaluation in Higher Education*, 40(5):712–724.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and 1 others. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

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.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Tobias van der Werff, Rik van Noord, and Antonio Toral. 2022. Automatic discrimination of human and neural machine translation: A study with multiple pretrained models and longer context. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 161–170, Ghent, Belgium. European Association for Machine Translation.

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

# A Appendix

## A.1 Attribution Score Analysis Methodology

This appendix details the mathematical formulation used to analyze attribution patterns between named entities (NEs) and non-NEs across our experimental conditions.

For each model, we analyze the distribution of attribution scores between NE and non-NE spans by computing their respective means. Specifically, we group the attribution scores based on whether each span is identified as an NE or not, then calculate the mean attribution score for each group. This yields two key metrics per model: the NE mean $\mu_{NE}$ (average attribution score for NE spans) and the non-NE mean $\mu_{non-NE}$ (average attribution score for non-NE spans).

To compare between NEs vs non-NEs, we first define a normalized difference score $\delta$ for each model:

$$\delta = \frac{\mu_{NE} - \mu_{non-NE}}{\mu_{non-NE}} \tag{1}$$

We then compute the difference in these normalized scores between experimental conditions:

For text modality experiments (standard vs. NE-masked fine-tuning):

$$\Delta_{text} = \delta_{standard} - \delta_{masked} \tag{2}$$

where $\delta_{standard}$ and $\delta_{masked}$ correspond to standard fine-tuning and NE-masked fine-tuning conditions respectively.

For speech modality experiments (unfrozen vs. frozen parameters):

$$\Delta_{speech} = \delta_{unfrozen} - \delta_{frozen} \tag{3}$$

where $\delta_{unfrozen}$ and $\delta_{frozen}$ correspond to unfrozen and frozen parameter settings respectively.

For bimodal comparison experiments:

$$\Delta_{bimodal} = \begin{cases} \delta_{text-unimodal} - \delta_{text-bimodal} \\ \delta_{speech-unimodal} - \delta_{speech-bimodal} \end{cases} \tag{4}$$

where $\delta_{text-unimodal}$ and $\delta_{speech-unimodal}$ are the $\delta$ values from the NE-masked text model and frozen speech model respectively, and $\delta_{text-bimodal}$ and $\delta_{speech-bimodal}$ are the corresponding $\delta$ values from the bimodal model.

These metrics allow us to quantify how different training approaches and modality combinations affect the model's reliance on named entities versus other textual elements.

## A.2 Complete Attribution Analysis Results

Tables 4, 5, 6 and 7 show the complete attribution results for the text-based classifiers (unimodal text modality), the speech-based classifiers (unimodal speech modality), the text-based classifiers (comparing the unimodal and bimodal settings) and the speech-based classifiers (comparing the unimodal and bimodal settings), respectively.

| Level | Method | ZH | | | EN | | |
|---|---|---|---|---|---|---|---|
| | | $\delta_{\text{std}}$ | $\delta_{\text{mask}}$ | $\Delta_{\text{text}}$ | $\delta_{\text{std}}$ | $\delta_{\text{mask}}$ | $\Delta_{\text{text}}$ |
| Word | m-mean | 0.13 | 0.06 | 0.07 | 0.90 | 0.34 | 0.56 |
| | m-L1 | 0.35 | 0.29 | 0.06 | 0.54 | 0.37 | 0.18 |
| | m-L2 | 0.33 | 0.26 | 0.07 | 0.48 | 0.31 | 0.16 |
| | s-max-m | 0.51 | 0.35 | 0.16 | 1.13 | 0.37 | 0.75 |
| | s-max-L1 | 0.51 | 0.35 | 0.16 | 1.11 | 0.38 | 0.72 |
| | s-max-L2 | 0.63 | 0.57 | 0.06 | 0.67 | 0.47 | 0.20 |
| Entity | m-mean | 0.23 | 0.09 | 0.14 | 1.28 | 0.47 | 0.81 |
| | m-L1 | 0.34 | 0.25 | 0.09 | 1.47 | 0.40 | 1.07 |
| | m-L2 | 0.30 | 0.21 | 0.09 | 0.52 | 0.31 | 0.21 |
| | s-max-m | 0.92 | 0.68 | 0.25 | 2.48 | 1.15 | 1.32 |
| | s-max-L1 | 0.93 | 0.68 | 0.25 | 2.45 | 1.17 | 1.29 |
| | s-max-L2 | 0.93 | 0.84 | 0.09 | 1.30 | 1.00 | 0.30 |

Table 4: Comparison of standard ($\delta_{\text{std}}$) and masked ($\delta_{\text{mask}}$) text-based classifiers' normalized difference scores, and their differences ($\Delta_{\text{text}}$) across different aggregation methods and span levels. (m = mean, s-max = signed-max.)

| Level | Method | ZH | | | EN | | |
|---|---|---|---|---|---|---|---|
| | | $\delta_u$ | $\delta_f$ | $\Delta_{\text{speech}}$ | $\delta_u$ | $\delta_f$ | $\Delta_{\text{speech}}$ |
| Word | mean | 3.84 | -0.11 | 3.95 | 0.78 | -0.59 | 1.37 |
| | L1 | -0.03 | -0.28 | 0.24 | 0.09 | -0.10 | 0.19 |
| | L2 | -0.04 | -0.28 | 0.23 | 0.09 | -0.10 | 0.19 |
| Entity | mean | 3.29 | -0.48 | 3.77 | 0.85 | -0.69 | 1.54 |
| | L1 | 0.01 | -0.28 | 0.29 | 0.16 | -0.07 | 0.23 |
| | L2 | 0.01 | -0.26 | 0.27 | 0.18 | -0.05 | 0.24 |

Table 5: Comparison of unfrozen ($\delta_u$) and frozen ($\delta_f$) speech-based classifiers' normalized difference scores, and their differences ($\Delta_{\text{speech}}$) across different aggregation methods and span levels.

| Level | Method | ZH | | | EN | | |
|---|---|---|---|---|---|---|---|
| | | $\delta_{\text{m-uni}}$ | $\delta_{\text{m-bi}}$ | $\Delta_{\text{t-bi}}$ | $\delta_{\text{m-uni}}$ | $\delta_{\text{m-bi}}$ | $\Delta_{\text{t-bi}}$ |
| Word | m-mean | 0.06 | -0.01 | 0.07 | 0.34 | 0.14 | 0.21 |
| | m-l1 | 0.29 | 0.26 | 0.03 | 0.37 | 0.34 | 0.02 |
| | m-l2 | 0.26 | 0.23 | 0.03 | 0.31 | 0.29 | 0.02 |
| | s-max-m | 0.35 | 0.27 | 0.08 | 0.37 | 0.23 | 0.14 |
| | s-max-l1 | 0.35 | 0.27 | 0.08 | 0.38 | 0.23 | 0.15 |
| | s-max-l2 | 0.57 | 0.57 | 0.01 | 0.47 | 0.46 | 0.01 |
| Entity | m-mean | 0.09 | 0.02 | 0.07 | 0.47 | 0.20 | 0.26 |
| | m-l1 | 0.25 | 0.21 | 0.04 | 0.40 | 0.36 | 0.04 |
| | m-l2 | 0.21 | 0.17 | 0.04 | 0.31 | 0.27 | 0.04 |
| | s-max-m | 0.68 | 0.57 | 0.11 | 1.15 | 0.88 | 0.27 |
| | s-max-l1 | 0.68 | 0.57 | 0.11 | 1.17 | 0.88 | 0.28 |
| | s-max-l2 | 0.84 | 0.83 | 0.004 | 1.00 | 0.98 | 0.03 |

Table 6: Comparison of unimodal ($\delta_{\text{m-uni}}$) and bimodal masked text modality's ($\delta_{\text{m-bi}}$) normalized difference scores, and their differences ($\Delta_{\text{t-bi}}$) across ZH and EN languages.

| Level | Method | ZH | | | EN | | |
|---|---|---|---|---|---|---|---|
| | | $\delta_{\text{uni}}$ | $\delta_{\text{bi}}$ | $\Delta_{\text{s-bi}}$ | $\delta_{\text{uni}}$ | $\delta_{\text{bi}}$ | $\Delta_{\text{s-bi}}$ |
| Word | mean | -0.11 | -0.44 | 0.33 | -0.59 | 0.67 | -1.26 |
| | L1 | -0.28 | -0.22 | -0.06 | -0.10 | 0.09 | -0.19 |
| | L2 | -0.28 | -0.23 | -0.05 | -0.10 | 0.11 | -0.21 |
| Entity | mean | -0.48 | -0.39 | -0.09 | -0.69 | 0.53 | -1.21 |
| | L1 | -0.28 | -0.21 | -0.07 | -0.07 | 0.11 | -0.17 |
| | L2 | -0.26 | -0.20 | -0.06 | -0.05 | 0.14 | -0.19 |

Table 7: Comparison of unimodal ($\delta_{\text{froz-uni}}$) and bimodal ($\delta_{\text{froz-bi}}$) frozen speech modality's normalized difference scores, and their differences ($\Delta_{\text{speech-bi}}$).