Machine Translation Models are Zero-Shot Detectors of Translation Direction

Michelle Wastl **Jannis Vamvas** **Rico Sennrich**
Department of Computational Linguistics, University of Zurich
{wastl,vamvas,sennrich}@cl.uzh.ch

Abstract

Detecting the translation direction of parallel text is useful not only for machine translation training and evaluation but also has forensic applications, such as resolving plagiarism or forgery allegations. In this work, we explore an unsupervised approach to translation direction detection based on the simple hypothesis that $p(\text{translation}|\text{original}) > p(\text{original}|\text{translation})$, motivated by the well-known simplification effect in translationese or machine-translationese. In experiments with multilingual machine translation models across 20 translation directions, we confirm the effectiveness of the approach for high-resource language pairs, achieving document-level accuracies of 82–96% for NMT-produced translations, and 60–81% for human translations, depending on the model used.[1]

1 Introduction

While the original translation direction of parallel text is often ignored or unknown in the machine translation community, research has shown that it can be relevant for training (Kurokawa et al., 2009; Ni et al., 2022) and evaluation (Graham et al., 2020).[2] Beyond machine translation, translation direction detection has practical applications in areas such as forensic linguistics, where determining the original of a document pair may help resolve plagiarism or forgery accusations.

Previous work has addressed translation (direction) detection with feature-based approaches, using features such as n-gram frequency statistics and POS tags for classification (Kurokawa et al., 2009; Volansky et al., 2013; Sominsky and Wintner, 2019) or unsupervised clustering (Nisioi, 2015; Rabinovich and Wintner, 2015). However, these methods require a substantial amount of text data, and cross-domain differences in the statistics used can overshadow differences between original and translationese text.

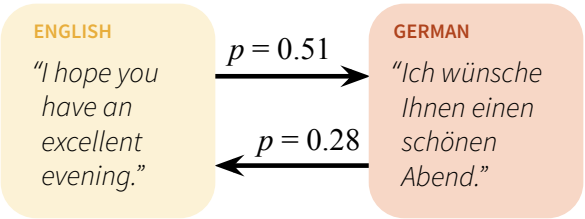

Figure 1: NMT models can be used for inferring the likely original translation direction of parallel text. In this example, the NMT model assigns a much higher probability to the German sentence given the English sentence than to the English sentence given the German sentence, indicating that the more likely original translation direction is English→German.

In this work, we explore the unsupervised detection of translation directions purely on the basis of a multilingual neural machine translation (NMT) system's translation probabilities. As illustrated in Figure 1, we hypothesize that $p(\text{translation}|\text{original}) > p(\text{original}|\text{translation})$, which, if it generally holds, would allow us to infer the original translation direction.

If the translation has been automatically generated, this hypothesis can be motivated by the fact that machine translation systems typically generate text with mode-seeking search algorithms, and consequently tend to over-produce high-frequency outputs and reduce lexical diversity (Vanmassenhove et al., 2019). However, even human translations are known for so-called translationese properties such as interference, normalization, and simplification, and a (relative) lack of lexical diversity (Teich, 2003; Volansky et al., 2013; Toral, 2019).

[1] Code and demo are available at https://github.com/ZurichNLP/translation-direction-detection

[2] As of today, training data is not typically filtered by translation direction, but we find evidence of a need for better detection in recent work. For example, Post and Junczys-Dowmunt (2023) show that back-translated data is more suited than crawled parallel data for document-level training, presumably because of translations in the crawled data that lack document-level consistency.

Findings of the Association for Computational Linguistics: ACL 2025, pages 1054–1074
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

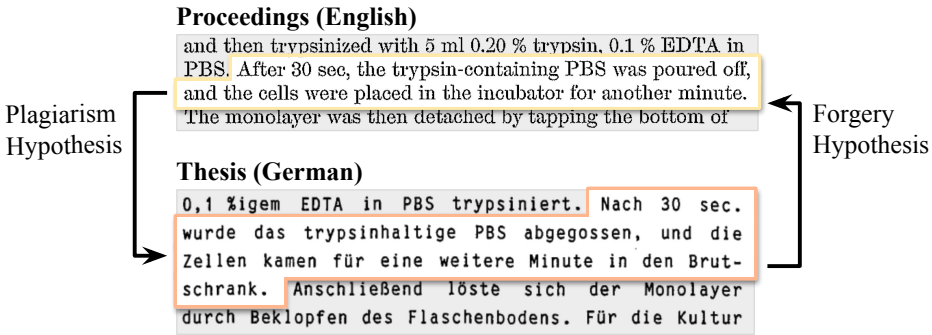

Figure 2: A recent forensic case in Germany underscores the relevance of translation direction detection (Ebbinghaus, 2022; Zenthöfer, 2022; Wikipedia, 2023). In 2022, two experts raised concerns about the originality of a German PhD thesis and suspected it to be plagiarized from a proceedings volume in English (*plagiarism hypothesis*). Further investigation showed, however, that the alleged English source could not be found in any library or database. This raised the possibility of a deliberate attempt to discredit the thesis author by fabricating the English book (*forgery hypothesis*). Initially, the debate focused on the dating of the typefaces and paper used to print the book, in addition to textual inconsistencies. A computational analysis of translation direction could provide additional evidence in this or similar cases. The illustration depicts one of the parallel passages identified by Weber (2022).

We test the approach on 20 translation directions, experimenting with 3 massively multilingual NMT models to predict the translation probabilities of human translations, NMT-produced translations, LLM-generated translations, and pre-neural machine translations. We find that the approach detects the translation direction of human translations with an average accuracy of 66% at the sentence level, and 80% for documents with ≥ 10 sentences. For the output of NMT systems, detection accuracy is even higher, but our hypothesis that $p(\text{translation}|\text{original}) > p(\text{original}|\text{translation})$ does not hold for the output of pre-neural systems.

To compare our unsupervised approach to a supervised baseline, we implement a modernized version of the approach proposed by Sominsky and Wintner (2019). A controlled comparison shows that a supervised approach can outperform our unsupervised one under ideal conditions (with in-domain labelled training data) for human translations, but performance drops in a cross-domain setting. Notably, for NMT-produced translations, the unsupervised approach remains competitive even when tested within the same domain.

Finally, we apply our method to a recent forensic case (Figure 2), where the translation direction of a German PhD thesis and an English book has been under dispute, finding additional evidence for the hypothesis that the English book is a forgery created to make the thesis appear plagiarized.

Our main contributions are the following:

- We propose a simple, unsupervised approach to translation direction detection based on the translation probabilities of NMT models.
- We demonstrate that the approach is effective for detecting the original translation direction of neural machine translations, and to a lesser extent, human translations in a variety of high-resource language pairs.
- We provide a qualitative analysis of detection performance and apply the method to a real-world forensic case.

2 Related Work

2.1 Translation (Direction) Detection

In an ideal scenario where large-scale annotated in-domain data is available, high accuracy can be achieved in translation direction detection at phrase and sentence level by training supervised systems based on various features such as word frequency statistics, POS n-grams or text embeddings (Sominsky and Wintner, 2019).

To reduce reliance on in-domain supervision, unsupervised methods that rely on clustering and consequent cluster labelling have also been explored for the related task of translationese detection (Rabinovich and Wintner, 2015; Nisioi, 2015). One could conceivably perform translation direction detection using similar methods, but this has the practical problem of requiring an expert for cluster labelling and poor open-domain performance. In a multi-domain scenario, Rabinovich and Wintner (2015) observe that clustering based on features

proposed by Volansky et al. (2013) results in clusters separated by domain rather than translation status. They address this by producing $2k$ clusters, k being the number of domains in their dataset, and labelling each. Clearly, labelling becomes more costly as the number of domains increases, which limits applicability to an open-domain scenario.

In contrast, we hypothesize that comparing translation probabilities remains a valid strategy across domains, and requires no resources other than NMT models that are competent for the respective language pair.

2.2 Translation Probabilities

Previous work has leveraged translation probabilities for tasks such as noisy parallel corpus filtering (Junczys-Dowmunt, 2018), machine translation evaluation (Thompson and Post, 2020), and paraphrase identification (Mallinson et al., 2017; Vamvas and Sennrich, 2022). Those approaches analyze translation probabilities symmetrically in two directions, which is also the case in this work.

3 Methods

Given a parallel sentence pair (x, y), the main task in this work is to identify the translation direction between a language X and a language Y, and, consequently, establish which side is the original and which is the translation. This is achieved by comparing the conditional translation probability $P(y|x)$ by an NMT model $M_{X\to Y}$ with the conditional translation probability $P(x|y)$ by a model $M_{Y\to X}$ operating in the inverse direction. Our core assumption is that segment pairs in the original translation direction are assigned higher conditional probabilities by NMT models than in the inverse direction, so if $P(y|x) > P(x|y)$, we predict that y is the translation, and x the original, and the original translation direction is $X \to Y$.

3.1 Detection on the Sentence Level

With a probabilistic autoregressive NMT model, we can obtain $P(y|x)$ as a product of the individual token probabilities:

$$P(y|x) = \prod_{j=1}^{|y|} p(y_j|y_{<j}, x) \tag{1}$$

We follow earlier work by Junczys-Dowmunt (2018); Thompson and Post (2020), and average token-level (log-)probabilities.[3]

$$P_{\text{tok}}(y|x) = P(y|x)^{\frac{1}{|y|}} \tag{2}$$

To detect the original translation direction (OTD), $P_{\text{tok}}(y|x)$ and $P_{\text{tok}}(x|y)$ are compared:

$$\text{OTD} = \begin{cases} X \to Y, & \text{if } P_{\text{tok}}(y|x) > P_{\text{tok}}(x|y) \\ Y \to X, & \text{otherwise} \end{cases}$$

3.2 Detection on the Document Level

We also study translation direction detection at the level of documents, as opposed to individual sentences. We assume that the sentences in the document are aligned 1:1, so that we can apply an NMT model trained at the sentence level to all n sentence pairs (x_i, y_i) in the document, and then aggregate the result.

Our approach is equivalent to the sentence-level approach in that we calculate the average token-level probability across the document, conditioned on the respective sentence in the other language:

$$P_{\text{tok}}(y|x) = [\prod_{i=1}^{n} \prod_{j=1}^{|y_i|} p(y_{i,j}|y_{i,<j}, x_i)]^{\frac{1}{|y_1|+\cdots+|y_n|}} \tag{3}$$

The criterion for the original translation direction is then again whether $P_{\text{tok}}(y|x) > P_{\text{tok}}(x|y)$.

3.3 On Directional Bias

A multilingual translation model (or a pair of bilingual models) may consistently assign higher probabilities in one translation direction than the other, thus biasing our prediction. This could be the result of training data imbalance, tokenization choices, or typological differences between the languages (Cotterell et al., 2018; Bugliarello et al., 2020).[4]

To allow for a cross-lingual comparison of bias despite varying data balance, we measure bias via the difference in accuracy between the two gold directions. An unbiased model should have similar accuracy in both. An extremely biased model that always predicts OTD $= X \to Y$ will achieve perfect accuracy on the gold direction $X \to Y$,

[3] The models we use have been trained with label smoothing (Szegedy et al., 2016), which has a cumulative effect on sequence-level probabilities (Yan et al., 2024). Averaging token-level probabilities can help mitigate this shortcoming.

[4] We note that Bugliarello et al. (2020) do not control for the original translation direction of their data. Re-examining their findings in view of our core hypothesis could be fruitful future work.

and zero accuracy on the reverse gold direction $Y \rightarrow X$. We will report the bias B as follows:

$$B = |acc(X \rightarrow Y) - acc(Y \rightarrow X)| \quad (4)$$

This yields a score that ranges from 0 (unbiased) to 1 (fully biased).

4 Experiments: Models and Data

4.1 Unsupervised Setting

We experiment with three multilingual machine translation models: M2M-100-418M (Fan et al., 2021), SMaLL-100 (Mohammadshahi et al., 2022), and NLLB-200-1.3B (NLLB Team et al., 2024).

The models are architecturally similar, all being based on the Transformer architecture (Vaswani et al., 2017), but they differ in the training data used, number of languages covered, model size, and, consequently, in translation quality. The comparison allows conclusions about how sensitive our method is to translation quality – NLLB-200-1.3B yields the highest translation quality of the three (Tiedemann and de Gibert, 2023), but we also highlight differences in data balance. English has traditionally been dominant in the amount of training data, and all three models aim to reduce this dominance in different ways, for example via large-scale back-translation (Sennrich et al., 2016) in M2M-100-418M and NLLB-200-1.3B. SMaLL-100 is a distilled version of the M2M-100-12B model, and samples training data uniformly across language pairs.

We test the approach on datasets from the WMT news/general translation tasks from WMT16 (Bojar et al., 2016), WMT22 (Kocmi et al., 2022), and WMT23 (Kocmi et al., 2023), which come annotated with document boundaries and the original language of each document. We include different years of WMT data not only to test the approach for different translation directions and translation types but also to rule out test set contamination for experiments with models, for which the test set predates the model release. The results for the individual WMT datasets are reported in Appendix F.

We also experiment with a subset of the FLORES-101 dataset (Goyal et al., 2022) to test the approach on indirect translations, where English was the original language of both sides of the parallel text. We divide the data into subsets based on several categorizations:

- **Translation direction**: the WMT data span 14 translation directions and 3 scripts (Latin, Cyrillic, Chinese).
- **Type of translation**: we distinguish between human translations (**HT**), which consist of (possibly multiple) reference translations, neural translation systems (**NMT**) (WMT 2016; 2022–2023; (incl. translations from machine translation systems specifically and, to a lesser extent, translations produced by large language models (LLMs)), and phrase-based or rule-based pre-neural systems from WMT 2016 (**pre-NMT**) as a third category.
- **Directness**: Given that the WMT data are *direct* translations from one side of the parallel text to the other, we perform additional experiments on translations for 6 FLORES language pairs (Bengali↔Hindi, Czech↔Ukrainian, German↔French, German↔Hindi, Chinese↔French, and Xhosa↔Zulu). This allows us to analyze the behavior of our approach on *indirect* sentence pairs where both sequences are translations from a third language (in this case, English).

We use HT and NMT translations from WMT16 as a validation set and the remaining translations for testing our approach. Table 1 shows test set statistics for our main experiments.

To demonstrate our approach on "real-world" data, we use text pertaining to a publicly documented plagiarism allegation case, in which translation-based plagiarism was the main focus (Ebbinghaus, 2022; Zenthöfer, 2022). We use 86 parallel segments in German and English from

	Source		Target Sentences		
Direction	Sents	Docs (≥10)	HT	NMT	Pre-NMT
cs→en	1448	129	2896	15928	16489
cs→uk	3947	112	3947	49381	-
de→en	1984	121	3968	17856	13491
de→fr	1984	73	1984	11904	-
en→cs	4111	204	6148	51480	27000
en→de	2037	125	4074	18333	18000
en→ru	4111	174	4111	47295	15000
en→uk	4111	174	4111	41147	-
en→zh	4111	204	6148	57591	-
fr→de	2006	71	2006	14042	-
ru→en	3739	136	3739	40836	13482
uk→cs	2812	43	2812	33744	-
uk→en	3844	88	3844	40266	-
zh→en	3851	162	5726	52140	-
Total	44096	1816	55514	491943	103462

Table 1: Statistics of the WMT data used in our main experiments. More granular statistics are provided in Appendix G.

LP		HT			NMT		
		→	←	Avg.	→	←	Avg.
en↔cs	supervised	64.70	**71.36**	**68.03**	65.43	71.98	68.71
	unsupervised	**68.85**	65.19	67.02	**71.87**	**78.30**	**75.09**
en↔ru	supervised	56.65	**81.74**	**69.19**	61.72	**82.97**	72.35
	unsupervised	**71.81**	54.05	62.93	**76.91**	71.98	**74.44**
en↔de	supervised	**87.05**	55.29	**71.17**	**89.52**	61.08	**75.30**
	unsupervised	56.38	**67.44**	61.91	62.69	**85.27**	73.98
Macro Avg.	supervised	**69.46**	**69.47**	**69.46**	**72.22**	72.01	72.12
	unsupervised	65.68	62.23	63.95	70.49	**78.51**	**74.50**

Table 2: Accuracy of the proposed unsupervised approach, using the M2M100 MT system, compared to the performance of XLM-R fine-tuned on the translation direction detection task with WMT data.

aligned excerpts of both the PhD thesis and the alleged source that were presented in a plagiarism analysis report (Weber, 2022). We extracted the segments with OCR and manually checked for OCR errors. An example translation from this dataset is given in Appendix I.

4.2 Supervised Baseline

In addition to the unsupervised approach, we fine-tune XLM-R (base) (Conneau et al., 2020) on the translation direction detection task to provide a supervised baseline for our experiments.

We extract the final hidden state for the first token by XLM-R for each segment of a translation, which serves as the feature representation for that segment. Inspired by the COMET architecture for MT evaluation (Rei et al., 2020), we then combine the resulting representations for each translation by concatenating the addition of the representations of the segments of a pair, their absolute difference, and their product.[5,6] The resulting concatenation is used as input to train the classification head.

For the training set, we take a sample of 1398–1400 source segments and their corresponding HT and NMT-based translations per direction from WMT16. Additionally, we sample another 100 source segments per direction and their corresponding HT, NMT and pre-NMT-based translations from WMT16 as a validation set. We then train bilingual classifiers for each en↔cs, en↔ru and en↔de. We validate settings with different learning rates and epochs (Appendix C) and report the results for the test set in Table 2 with the setting that achieves the highest accuracies for the corresponding language pair[7] on the validation set. The test set consists of HT and NMT sentence pairs from WMT21, WMT22 and WMT23, which is equivalent to the unsupervised experiments apart from the WMT16.

Since training the supervised systems on larger datasets could improve their performance, we experiment with training them on subsets of the Europarl corpus (Koehn, 2005; Ustaszewski, 2019). Although higher accuracies are reached when the systems are tested on in-domain data, their performance drops below the WMT-based system when tested on out-of-domain data (WMT). Hence, we choose to describe the WMT-based model, which performs best on WMT data in the main part of this paper (Subsection 5.1) and give a detailed description and results of the Europarl-based system in Appendix E.

5 Results

5.1 Sentence-level Classification

The sentence-level results are shown in Tables 3 (HT), 4 (NMT), and 5 (pre-NMT).

Table 3 compares the results for HT across all models. As a general result, we find that it is not NLLB, but M2M-100 that on average yields the best results for HT on the sentence level, with

[5]We deviate slightly from COMET by using the addition of the two segment representations instead of concatenating them sequentially, to avoid introducing input order bias.

[6]Furthermore, we experiment with encoding the source segment jointly with the translation. This yields similar results, but introduces input order bias. The test set results for this alternative architecture are listed in Appendix D.

[7]en↔cs/ru: learning rate 1e-05, epoch 2; en↔de: learning rate 1e-05, epoch 4.

	M2M-100-418M			SMaLL-100			NLLB-200-1.3B		
Language Pair	→	←	Avg.	→	←	Avg.	→	←	Avg.
HT en↔cs	68.85	65.19	**67.02**	63.08	69.37	66.22	54.05	68.78	61.42
HT en↔de	56.38	67.44	**61.91**	58.62	63.10	60.86	59.70	47.76	53.73
HT en↔ru	71.81	54.05	62.93	68.38	57.56	**62.97**	67.40	49.08	58.24
HT en↔uk	71.95	69.56	**70.76**	70.49	68.83	69.66	47.21	64.00	55.61
HT en↔zh	54.25	84.30	**69.27**	56.41	80.54	68.48	17.81	82.52	50.16
HT cs↔uk	52.44	74.40	63.42	59.26	70.52	**64.89**	47.68	76.67	62.18
HT de↔fr	89.72	50.50	70.11	85.48	57.68	71.58	86.29	62.16	**74.23**
Macro-Avg.	66.49	66.49	**66.49**	65.96	66.80	66.38	54.31	64.42	59.37

Table 3: Accuracy of three different models when detecting the direction for human translation. The first column per model reports accuracy for sentence pairs with left-to-right gold direction (e.g., en→cs), the second column for sentence pairs with the reverse gold direction (e.g., en←cs). The last column reports the macro-average across both directions. The best average result for each language pair is printed in bold.

Language Pair	→	←	Avg.
NMT en↔cs	71.87	78.30	75.09
NMT en↔de	62.69	85.27	73.98
NMT en↔ru	76.91	71.98	74.44
NMT en↔uk	75.01	79.31	77.16
NMT en↔zh	64.29	90.29	77.29
NMT cs↔uk	72.83	79.15	75.99
NMT de↔fr	90.65	51.60	71.13
Macro-Avg.	73.46	76.56	75.01

Table 4: Accuracy of M2M-100 when detecting the translation direction of NMT-produced translations.

Language Pair	→	←	Avg.
Pre-NMT en↔cs	41.97	42.59	42.28
Pre-NMT en↔de	33.33	54.30	43.81
Pre-NMT en↔ru	37.98	39.01	38.49
Macro-Avg.	37.76	45.30	41.53

Table 5: Accuracy of M2M-100 when detecting the translation direction of sentences translated with pre-NMT systems.

Language Pair	→	←	Avg.
en↔ru	75.75	68.08	71.91
en↔uk	72.76	72.56	72.66
en↔zh	58.82	90.54	74.68
Macro-Avg.	69.11	77.06	73.08

Table 6: Accuracy on translations generated by LLMs with M2M-100.

SMaLL-100 a close second. Hence, we report results for experiments with M2M-100 in the following, while performance of the other models is reported in Appendix A and B.

A second result is that the translation detection works best for NMT-based translations (75.0% macro-average), second-best for HT (66.5% macro-average), and worst for pre-NMT (41.5% macro-average). The fact that performance for pre-neural systems is below chance level indicates that the NMT systems we use tend to assign low probabilities to the (often ungrammatical) outputs of pre-neural systems.

A third result is that accuracy varies by language pair. Among the language pairs tested, accuracy of M2M-100 ranges from 61.9% (en↔de) to 70.8% (en↔uk) for HT, and from 71.1% (de↔fr) to 77.3% (en↔zh) for NMT.

In comparison to the results of the supervised system, we show in Table 2 that while our unsupervised approach is outperformed on HT, it is competitive for NMT-based translation. Taking into consideration the (cross-domain) results of the Europarl-based models shown in Table 17 as well, the main benefits of the unsupervised approach are highlighted: independence from training data and flexibility across languages and domains.

5.2 LLM-generated Translations

The main focus of this paper lies on three different translation types: HT, NMT, and pre-NMT. However, LLMs have also shown strong translation capabilities (Kocmi et al., 2023). Hence, GPT-4 was considered as one of the translation systems in the

Language Pair	Ratio of →	Ratio of ←
HT bn↔hi	66.80 %	33.20 %
HT de↔hi	48.72 %	51.28 %
HT cs↔uk	42.69 %	57.31 %
HT de↔fr	84.78 %	15.22 %
HT zh↔fr	91.70 %	8.30 %
HT xh↔zu	45.06 %	54.94 %

Table 7: Percentage of predictions by M2M-100 for each translation direction when neither is the true translation direction (English-original FLORES).

WMT23 shared task and its outputs are therefore part of the NMT test set in this work. Table 6 shows the results of our translation direction detection approach on the GPT-4-generated test subset in isolation to document its performance on this additional translation type. For all three language pairs, our approach reaches accuracies that are comparable to the ones reached for the same pairs' NMT translations.

5.3 Indirect Translations

With an experiment on the English-original FLORES data, we evaluate our approach on the special case that neither side is the original. As shown in Table 7, our approach yields relatively balanced predictions on human translations for cs↔uk and xh↔zu, predicting each direction a roughly equal number of times. For de↔fr, we again find that the model predicts de→fr more frequently than the reverse direction. The result for fr↔zh display an even larger degree of disparity.

5.4 Directional Bias

When analyzing Table 3 for directional bias, we observe that M2M-100 is especially biased in the directions de→fr ($B = 0.39$) and zh→en ($B = 0.30$). While we expected a general bias towards x→en due to the dominance of English in training data, we find that the direction and strength of the bias vary across language pairs and models. An extreme result is NLLB for en↔zh, with $B = 0.64$ towards zh→en.

There appears to be an inclination for bias towards languages closely related to English for some language pairs, such as de↔fr and zh↔fr (Table 7) in contrast to language pairs, with both languages being closely related to (and including) English, such as en↔fr. The discrepancy in balance between de↔hi and zh↔fr indicates, however, that relatedness to English might not have a strong general influence on our approach, but rather that there is a bias for →French and Chinese→ when applying it with M2M-100 or SMaLL-100.

Furthermore, the supervised systems exhibit higher bias (Table 2) than the unsupervised approach for all language pairs and translation types, indicating that directional bias is an issue for both supervised and unsupervised approaches.

We leave it to future work to explore whether bias can be reduced via different normalization strategies, a language pair specific bias correction term, or different model training. At present, our recommendation is to be mindful in the choice of NMT model and to perform validation before trusting the results of a previously untested NMT model for translation direction detection.

5.5 Qualitative Analysis

A qualitative comparison of sources and translations, as illustrated in Table 8, reveals that factors such as normalization, simplification, word order interference, and sentence length influence the detection of translation direction. In Example 1, an English HT translates the German source fully, while the NMT omits half of the content, showing a high degree of simplification. Our method recognizes the simplified NMT version as a translation, but not the more complete HT. Example 2 demonstrates normalization to a high degree. This shows that NMT-models assign extremely low probabilities when the target is more ungrammatical than the source. The third example indicates that translations exhibiting normalization, simplification, and interference to a higher degree are more likely to be identified. In Example 4, source language interference in terms of word order and choice significantly impacts the detection; the more literal translation mirroring the source's word order is recognized, while the more liberal translation is not. Finally, Example 5 highlights challenges with short sentences: The German phrase *Mit freundlichen Grüßen* is fairly standardized, while its French equivalents can vary in use and context, adding to the ambiguity and affecting the probability distribution in NMT. Hence, our approach fails to identify any of the French translations without additional context.

Furthermore, Table 8 shows that our approach more easily identifies translations that are simpler in terms of verbosity and sentence complexity, e.g.: Examples 2, 3, 4. While previous research indicates that verbosity is not the most reliable feature

	Sentences	→	←	Rel. Difference
1	*DE: Mit dem Programm "Guten Tag, liebes Glück" ist er seit 2020 auf Tour.*			
	EN: He has been on tour with the programme "Guten Tag, liebes Glück" since 2020. (HT)	0.145	0.558	0.26
	EN: He has been on tour since 2020. (NMT)	0.272	0.092	**2.95**
2	*EN: please try to perfprm thsi procedures"*			
	DE: bitte versuchen Sie es mit diesen Verfahren (HT)	0.246	0.010	**24.29**
	DE: Bitte versuchen Sie, diese Prozeduren durchzuführen" (NMT)	0.586	0.025	**23.59**
3	*EN: If costs for your country are not listed, please contact us for a quote.*			
	DE: Wenn die Kosten für Ihr Land nicht aufgeführt sind, wenden Sie sich für einen Kostenvoranschlag an uns. (HT)	0.405	0.525	0.77
	DE: Wenn die Kosten für Ihr Land nicht aufgeführt sind, kontaktieren Sie uns bitte für ein Angebot. (NMT)	0.697	0.585	**1.19**
4	*EN: Needless to say, it was chaos.*			
	DE: Es war natürlich ein Chaos. (HT)	0.119	0.372	0.32
	DE: Unnötig zu sagen, es war Chaos. (NMT)	0.755	0.591	**1.28**
5	*DE: Mit freundlichen Grüßen*			
	FR: Cordialement (HT)	0.026	0.107	0.24
	FR: Sincèrement (NMT)	0.015	0.083	0.18
	FR: Sincères amitiés (NMT)	0.062	0.160	0.39
	FR: Avec mes meilleures salutations (NMT)	0.215	0.353	0.61

Table 8: Qualitative comparison of sentence pairs. Source sentences are marked in *italics*, and gold direction is always →. Relative probability difference > 1 indicates that translation direction was successfully identified, and is highlighted in bold. The probabilities are generated by M2M-100.

Language Pair	→	←	Avg.
HT en↔cs	88.24	80.62	84.43
HT en↔de	70.40	88.43	79.41
HT en↔ru	96.55	54.41	75.48
HT en↔zh	67.65	97.53	82.59
Macro-Avg.	80.71	80.25	80.48

Table 9: Document-level classification: Accuracy of M2M-100 when detecting the translation direction of human translations at the document level (documents with ≥ 10 sentences).

Language Pair	→	←	Avg.
NMT en↔cs	96.78	99.27	98.03
NMT en↔de	91.06	99.18	95.12
NMT en↔ru	98.39	94.60	96.50
NMT en↔zh	86.33	98.62	92.47
Macro-Avg.	93.14	97.92	95.53

Table 10: Document-level classification: Accuracy of M2M-100 when detecting the translation direction of NMT translations at the document level (documents with ≥ 10 sentences).

for supervised approaches (Sominsky and Wintner, 2019), it is likely a helpful feature for the unsupervised approach.

Misclassified short examples as in Example 5 are not a rarity in our experiments. Our findings show that an average accuracy comparable to that reported in Table 3 is attained starting at a sentence length between 60 and 70 characters.[8] Additionally, we observed a trend where the accuracy of direction detection increases as the length of the sentences/documents grows. This aligns with previous unsupervised approaches, which also documented higher accuracy the larger the text chunks that were used, although there, reliable results were reported on a more extreme scale starting from text chunks with a length of 250 tokens (Rabinovich and Wintner, 2015).

5.6 Document-Level Classification

Accuracy scores for document-level results by M2M-100 (best-performing system at sentence level) are presented in Tables 9 (HT) and 10 (NMT). We consider documents with at least 10 sentences, and language pairs with at least 100 such documents in both directions.

The table shows that the sentence-level results

[8] We used SMaLL-100 for this analysis. See Appendix H.

are amplified at the document level. Translation direction detection accuracy for human translations reaches a macro-average of 80.5%, while the document-level accuracy for translations generated by NMT systems reaches 95.5% on average.

5.7 Application to Real-World Forensic Case

Finally, we apply our approach to the 86 segment pairs of the plagiarism allegation case. We treat the segments as a single document and classify them with M2M-100 using the document-level approach defined in Section 3.2. We find that, according to the model, it is more probable that the English segments are translations of the German segments than vice versa.

We validate our analysis using a permutation test. The null hypothesis is that the model probabilities for both potential translation directions are drawn from the same distribution. In order to perform the permutation test, we swap the segment-level probabilities $P(y_i|x_i)$ and $P(x_i|y_i)$ for randomly selected segments i before calculating the difference between the document-level probabilities $P(y|x)$ and $P(x|y)$. We repeat this process 10,000 times and calculate the p-value as twice the proportion of permutations that yield a difference at least as extreme as the observed difference. Obtaining a p-value of 0.0002, we reject the null hypothesis and conclude that our approach makes a statistically significant prediction that the English segments are translated from the German segments.

Overall, our analysis supports the hypothesis that German is indeed the language of origin in this real-world dataset (*forgery hypothesis*; Figure 2). Nevertheless, we recommend that additional evidence of different approaches (automated, manual, and qualitative) should be considered before drawing a final conclusion, given the error rate of 5–21% that we observed in experiments on WMT (en↔de).

6 Conclusion

We proposed a novel approach to detecting the translation direction of parallel texts, using only an off-the-shelf multilingual NMT system. Experiments on WMT data showed that our approach, without any task-specific supervision, is able to detect the translation direction of NMT-produced translations with relatively high accuracy, proving competitive to supervised baselines. Accuracy increases to 96% if the classifier is provided with at least 10 sentences per document. We also found a robust accuracy for translations by human translators. Finally, we applied our approach to a real-world forensic case and found that it supports the hypothesis that the English book is a forgery. Future work should explore whether our approach can be improved by mitigating directional bias of the NMT model used. Another open question is to what degree our approach will generalize to document-level translation.

Limitations

While the proposed approach is simple and effective, there are some limitations that might make its application more difficult in practice:

Sentence alignment: We performed our experiments on sentence-aligned parallel data, where each sentence in one language has a corresponding sentence in the other language. In practice, parallel documents might have one-to-many or many-to-many alignments, which would require custom preprocessing or the use of models that can directly estimate document-level probabilities.

Translation production workflow: Our main experiments used academic data from the WMT translation task, where care is taken to ensure that different translation methods are clearly separated: NMT translations did not undergo human post-editing, and human translators were instructed to work from scratch. In practice, parallel documents might have undergone a mixture of translation strategies, which makes it more difficult to predict the accuracy of our approach. Specifically, we found that our approach has less-than-chance accuracy on pre-NMT translations. Applying our approach to web-scale parallel corpus filtering (Kreutzer et al., 2022; Ranathunga et al., 2024) might therefore require additional filtering steps to exclude translations of lower quality.

Low-resource languages: Our experiments required test data for both translation directions, which limited the set of languages we could test. While the community has created reference translations for many low-resource languages, the translation directions are usually not covered symmetrically. For example, the test set of FLORES (Goyal et al., 2022) has been translated from English into many languages, but not vice versa. Thus, apart from Table 7, we have not tested our approach on low-resource languages, and it is possible that

the accuracy of our approach is lower for such languages, in parallel with the lower translation quality of NMT models for low-resource languages.

Ethical Considerations

Translation direction detection has a potential application in forensic linguistics, where reliable accuracy is crucial. Our experiments show that accuracy can vary depending on the language pair, the NMT model used for detection, as well as the translation strategy and the length of the input text. Before our approach is applied in a forensic setting, we recommend that its accuracy be validated in the context of the specific use case.

In Section 5.7, we tested our approach on a real-world instance of such a case, where one party has been accused of plagiarism, but the purported original is now suspected to be a forgery. This case is publicly documented and has been widely discussed in German-speaking media (e.g., Ebbinghaus 2022; Zenthöfer 2022; Wikipedia 2023). For this experiment, we used 86 sentence pairs from the two (publicly available) books that are the subject of this case. However, the case has not been definitively resolved, as legal proceedings are still ongoing. No author of this paper is involved in the legal proceedings. We therefore refrain from publicly releasing the dataset of sentence pairs we used for this experiment.

Acknowledgements

We thank the anonymous reviewers for their helpful comments and suggestions. This project was funded by the Swiss National Science Foundation (project MUTAMUR; no. 213976 and project InvestigaDiff; no. 10000503).

References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649, Online. Association for Computational Linguistics.

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.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Uwe Ebbinghaus. 2022. Geschichte eines Vernichtungsversuchs. *Frankfurter Allgemeine Zeitung*.

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 Auliy, and Armand Jouliny. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22:1–38.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow,

Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X: Papers*, pages 79–86.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

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

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. SMaLL-100: Introducing Shallow Multilingual Machine Translation Model for Low-Resource Languages. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 8348–8359.

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

Sergiu Nisioi. 2015. Unsupervised classification of translated texts. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9103:323–334.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation.

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

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Surangika Ranathunga, Nisansa De Silva, Velayuthan Menan, Aloka Fernando, and Charitha Rathnayake. 2024. Quality does matter: A detailed look at the quality and utility of web-mined parallel corpora. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 860–880, St. Julian's, Malta. Association for Computational Linguistics.

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

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ilia Sominsky and Shuly Wintner. 2019. Automatic detection of translation direction. *International Conference Recent Advances in Natural Language Processing, RANLP*, 2019-Septe:1131–1140.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Elke Teich. 2003. *A Methodology for the Investigation of Translations and Comparable Texts*. De Gruyter Mouton, Berlin, Boston.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Jörg Tiedemann and Ona de Gibert. 2023. The OPUS-MT dashboard – a toolkit for a systematic evaluation of open machine translation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 315–327, Toronto, Canada. Association for Computational Linguistics.

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

Michael Ustaszewski. 2019. Optimising the Europarl corpus for translation studies with the EuroparlExtract toolkit. *Perspectives*, 27(1):107–123.

Jannis Vamvas and Rico Sennrich. 2022. NMTScore: A multilingual analysis of translation-based text similarity measures. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 198–213, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Stefan Weber. 2022. Gutachten zur Einhaltung der Regeln guten wissenschaftlichen Arbeitens in der Dissertation „Untersuchung zur Chemotaxis von Fibrosarkomzellen in vitro“ Universität Hamburg, 1987. Technical report, Salzburg.

Wikipedia. 2023. Colchicine – 100 years of research — Wikipedia, die freie Enzyklopädie.

Jianhao Yan, Jin Xu, Fandong Meng, Jie Zhou, and Yue Zhang. 2024. DC-MBR: Distributional cooling for minimum Bayesian risk decoding. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4423–4437, Torino, Italia. ELRA and ICCL.

Jochen Zenthöfer. 2022. Chronik einer Plagiats-Intrige. *Frankfurter Allgemeine Zeitung*.

A Comparison of Models (Sentence Level)

	M2M-100-418M			SMaLL-100			NLLB-200-1.3B		
Language Pair	→	←	Avg.	→	←	Avg.	→	←	Avg.
NMT en↔cs	71.87	78.30	**75.09**	67.96	79.11	73.53	62.42	77.89	70.16
NMT en↔de	62.69	85.27	73.98	67.05	80.96	**74.00**	73.51	74.36	73.93
NMT en↔ru	76.91	71.98	**74.44**	74.34	72.30	73.32	78.87	57.15	68.01
NMT en↔uk	75.01	79.31	**77.16**	73.74	78.27	76.01	58.48	79.56	69.02
NMT en↔zh	64.29	90.29	**77.29**	66.54	87.62	77.08	25.42	90.54	57.98
NMT cs↔uk	72.83	79.15	75.99	77.33	76.04	**76.68**	70.55	76.59	73.57
NMT de↔fr	90.65	51.60	71.13	86.83	59.19	**73.01**	79.44	57.58	68.51
Macro-Avg.	73.46	76.56	**75.01**	73.40	76.21	74.81	64.10	73.38	68.74

Table 11: Accuracy of three different models when detecting the translation direction of NMT-produced translations. The first column reports accuracy for sentence pairs with left-to-right gold direction (e.g., en→cs), the second column for sentence pairs with the reverse gold direction (e.g., en←cs). The last column reports the macro-average across both directions. The best result for each language pair is printed in bold.

	M2M-100-418M			SMaLL-100			NLLB-200-1.3B		
Language Pair	→	←	Avg.	→	←	Avg.	→	←	Avg.
Pre-NMT en↔cs	41.97	42.59	**42.28**	37.88	45.42	41.65	16.36	35.34	25.85
Pre-NMT en↔de	33.33	54.30	**43.81**	36.18	48.57	42.37	18.73	26.20	22.47
Pre-NMT en↔ru	37.98	39.01	**38.49**	35.71	39.19	37.45	19.71	16.04	17.88
Macro-Avg.	37.76	45.30	**41.53**	36.59	44.39	40.49	18.27	25.86	22.07

Table 12: Accuracy of three different models when detecting the translation direction of sentences translated with pre-NMT systems. The best result for each language pair is printed in bold.

B Comparison of Models (Document Level)

	M2M-100-418M			SMaLL-100			NLLB-200-1.3B		
Language Pair	→	←	Avg.	→	←	Avg.	→	←	Avg.
HT en↔cs	88.24	80.62	**84.43**	78.92	89.15	84.03	55.88	86.05	70.96
HT en↔de	70.40	88.43	**79.41**	73.60	82.64	78.12	68.00	45.45	56.73
HT en↔ru	96.55	54.41	75.48	95.40	61.03	**78.22**	82.18	39.71	60.94
HT en↔zh	67.65	97.53	82.59	71.57	96.30	**83.93**	3.92	96.91	50.42
Macro-Avg.	80.71	80.25	80.48	79.87	82.28	**81.08**	52.50	67.03	59.76

Table 13: Accuracy of three different models when detecting the translation direction of human-translated documents. The best result for each language pair is printed in bold.

	M2M-100-418M			SMaLL-100			NLLB-200-1.3B		
Language Pair	→	←	Avg.	→	←	Avg.	→	←	Avg.
NMT en↔cs	96.78	99.27	**98.03**	94.64	97.81	96.22	86.06	95.62	90.84
NMT en↔de	91.06	99.18	95.12	93.85	97.12	95.49	96.65	95.06	**95.85**
NMT en↔ru	98.39	94.60	**96.50**	97.05	95.12	96.08	99.20	72.75	85.97
NMT en↔zh	86.33	98.62	92.47	90.62	98.16	**94.39**	13.14	98.39	55.76
Macro-Avg.	93.14	97.92	95.53	94.04	97.05	**95.55**	73.76	90.46	82.11

Table 14: Accuracy of three different models when detecting the translation direction of documents translated with NMT systems. The best result for each language pair is printed in bold.

C Supervised Baseline (WMT) Result Validation Set

	HT			NMT			Pre-NMT			
Language Pair	→	←	Avg.	→	←	Avg.	→	←	Avg.	Overall Avg.
Learning Rate: 1e-05, Epoch 2										
en↔cs	83.00	89.00	86.00	87.00	89.00	88.00	89.00	89.00	89.00	**87.67**
en↔de	84.00	78.00	81.00	85.00	78.00	81.50	77.00	82.00	79.50	80.67
en↔ru	80.00	96.00	88.00	90.00	96.00	93.00	94.00	84.00	89.00	**90.00**
Macro-Avg.	82.33	87.67	85.00	87.33	87.67	87.50	86.67	85.00	85.83	86.11
Learning Rate: 1e-05, Epoch 4										
en↔cs	84.00	88.00	86.00	90.00	86.00	88.00	85.00	92.00	88.50	87.50
en↔de	92.00	73.00	82.50	93.00	74.00	83.50	71.00	93.00	82.00	**82.67**
en↔ru	87.00	87.00	87.00	94.00	84.00	89.00	87.00	92.00	89.50	88.50
Macro-Avg.	87.67	82.67	85.17	92.33	81.33	86.83	81.00	92.33	86.67	**86.22**
Learning Rate: 1e-05, Epoch 5										
en↔cs	85.00	87.00	86.00	90.00	85.00	87.50	84.00	92.00	88.00	87.17
en↔de	85.00	79.00	82.00	89.00	77.00	83.00	77.00	86.00	81.50	82.17
en↔ru	75.00	97.00	86.00	86.00	95.00	90.50	94.00	79.00	86.50	87.67
Macro-Avg.	81.67	87.67	84.67	88.33	85.67	87.00	85.00	85.67	85.33	85.67
Learning Rate: 2e-05, Epoch 2										
en↔cs	81.00	88.00	84.50	82.00	87.00	84.50	87.00	84.00	85.50	84.83
en↔de	67.00	88.00	77.50	68.00	90.00	79.00	89.00	67.00	78.00	78.17
en↔ru	66.00	96.00	81.00	74.00	97.00	85.50	96.00	71.00	83.50	83.33
Macro-Avg.	71.33	90.67	81.00	74.67	91.33	83.00	90.67	74.00	82.33	82.11
Learning Rate: 2e-05, Epoch 4										
en↔cs	82.00	85.00	83.50	84.00	87.00	85.50	83.00	85.00	84.00	84.33
en↔de	86.00	71.00	78.50	89.00	74.00	81.50	72.00	85.00	78.50	79.50
en↔ru	85.00	84.00	84.50	87.00	84.00	85.50	85.00	83.00	84.00	84.67
Macro-Avg.	84.33	80.00	82.17	86.67	81.67	84.17	80.00	84.33	82.17	82.83
Learning Rate: 2e-05, Epoch 5										
en↔cs	91.00	84.00	87.50	89.00	80.00	84.50	82.00	91.00	86.50	86.17
en↔de	83.00	76.00	79.50	87.00	78.00	82.50	78.00	84.00	81.00	81.00
en↔ru	83.00	85.00	84.00	85.00	88.00	86.50	86.00	85.00	85.50	85.33
Macro-Avg.	85.67	81.67	83.67	87.00	82.00	84.50	82.00	86.67	84.33	84.17
Learning Rate: 3e-05, Epoch 2										
en↔cs	81.00	91.00	86.00	81.00	89.00	85.00	89.00	82.00	85.50	85.50
en↔de	0.00	100.00	50.00	0.00	100.00	50.00	100.00	0.00	50.00	50.00
en↔ru	59.00	98.00	78.50	61.00	99.00	80.00	98.00	55.00	76.50	78.33
Macro-Avg.	46.67	96.33	71.50	47.33	96.00	71.67	95.67	45.67	70.67	66.83
Learning Rate: 3e-05, Epoch 4										
en↔cs	77.00	92.00	84.50	80.00	91.00	85.50	88.00	80.00	84.00	84.67
en↔de	100.00	0.00	50.00	100.00	0.00	50.00	0.00	100.00	50.00	50.00
en↔ru	63.00	98.00	80.50	72.00	98.00	85.00	98.00	65.00	81.50	82.33
Macro-Avg.	80.00	63.33	71.67	84.00	63.00	73.50	62.00	81.67	71.83	72.33
Learning Rate: 3e-05, Epoch 5										
en↔cs	75.00	91.00	83.00	76.00	88.00	82.00	90.00	75.00	82.50	82.50
en↔de	100.00	0.00	50.00	100.00	0.00	50.00	0.00	100.00	50.00	38.89
en↔ru	69.00	98.00	83.50	75.00	96.00	85.50	95.00	71.00	83.00	84.00
Macro-Avg.	81.33	63.00	72.17	83.67	61.33	72.50	61.67	82.00	71.83	72.17
Learning Rate: Dynamic, Epoch 2										
en↔cs	83.00	88.00	85.50	80.00	85.00	82.50	88.00	82.00	85.00	84.33
en↔de	81.00	83.00	82.00	79.00	83.00	81.00	85.00	79.00	82.00	81.67
en↔ru	83.00	90.00	86.50	89.00	90.00	89.50	90.00	82.00	86.00	87.33
Macro-Avg.	82.33	87.00	84.67	82.67	86.00	84.33	87.67	81.00	84.33	84.44
Learning Rate: Dynamic, Epoch 4										
en↔cs	83.00	89.00	86.00	85.00	90.00	87.50	88.00	83.00	85.50	86.33
en↔de	88.00	75.00	81.50	93.00	78.00	85.50	78.00	83.00	80.50	82.50
en↔ru	79.00	88.00	83.50	87.00	88.00	87.50	90.00	81.00	85.50	85.50
Macro-Avg.	83.33	84.00	83.67	88.33	85.33	86.83	85.33	82.33	83.83	84.78
Learning Rate: Dynamic, Epoch 5										
en↔cs	75.00	92.00	83.50	77.00	90.00	83.50	90.00	80.00	85.00	84.00
en↔de	88.00	77.00	82.50	88.00	78.00	83.00	77.00	82.00	79.50	81.67
en↔ru	79.00	91.00	85.00	87.00	88.00	87.50	89.00	79.00	84.00	85.50
Macro-Avg.	80.67	86.67	83.67	84.00	85.33	84.67	85.33	80.33	82.83	83.72

Table 15: Accuracy of the supervised baseline when detecting the translation direction of HT, NMT, and Pre-NMT-produced translations across various checkpoints and learning rates on the WMT validation set.

D Supervised Baseline (WMT) Result Validation Set with Joint Encoding

LP	Input Order	HT			NMT		
		→	←	Avg.	→	←	Avg.
en↔de	source + translation	79.51	65.45	72.48	83.13	69.23	76.18
	translation + source	78.59	60.86	69.72	83.91	65.57	74.74
en↔ru	source + translation	66.13	80.73	73.43	70.31	81.88	76.1
	translation + source	67.64	83.24	75.44	69.57	84.05	76.81
en↔cs	source + translation	61.26	69.3	65.28	62.24	69.65	65.94
	translation + source	61.7	66.89	64.3	63.15	67.51	65.33

Table 16: WMT test set results of the joint encoding approach, where the translated segment is appended to the source and vice versa.

Model and Data

In addition to the siamese architecture outlined in Section 4.2, which encodes each text segment independently before concatenating their respective embeddings, we also investigate an alternative approach that involves joint encoding of the source and translation texts. Specifically, the two text segments are concatenated with a special delimiter token ("</s>") inserted between them and then passed together through the encoder (Ranasinghe et al., 2020; Rei et al., 2022). The joint encoding system is trained using the same hyperparameters as outlined in Section 4.2, ensuring a consistent comparison.

To mitigate potential input order bias, we reversed the input order for half of the training data. However, a comparison of the results for the WMT test set with differing input order of the source and the target (Table 16) shows that this strategy did not eliminate the bias.

Ultimately, we observed that the siamese architecture demonstrated greater stability with respect to input order, without incurring a significant performance trade-off. Based on these findings, we focus on the siamese architecture as the main baseline presented in this paper.

E Supervised Baseline (Europarl) and Cross-Domain Application

LP	Approach	HT (Europarl)			HT (WMT)			NMT (WMT)		
		→	←	Avg.	→	←	Avg.	→	←	Avg.
en↔cs	sup (WMT)	—	—	—	64.70	**71.36**	**68.03**	65.43	71.98	68.71
	sup (Europarl)	71.51	83.14	77.32	**82.03**	48.39	65.21	**82.72**	50.18	66.45
	unsup	—	—	—	68.85	65.19	67.02	71.87	**78.30**	**75.09**
en↔de	sup (WMT)	—	—	—	**87.05**	55.29	**71.17**	**89.52**	61.08	**75.30**
	sup (Europarl)	82.12	70.80	76.46	80.97	45.36	63.17	84.15	39.70	61.93
	unsup	—	—	—	56.38	**67.44**	61.91	62.69	**85.27**	73.98
en↔ru	sup (WMT)	—	—	—	56.65	**81.74**	**69.19**	61.72	**82.97**	72.35
	sup (Europarl)	—	—	—	—	—	—	—	—	—
	unsup	—	—	—	**71.81**	54.05	62.93	**76.91**	71.98	**74.44**
de↔fr	sup (WMT)	—	—	—	—	—	—	—	—	—
	sup (Europarl)	59.80	56.67	58.24	58.28	52.78	55.53	58.85	53.38	56.11
	unsup	—	—	—	—	—	—	—	—	—

Table 17: Comparison of supervised and unsupervised approaches on the translation direction detection task for HT and NMT datasets from WMT and Europarl. The supervised approach involves fine-tuning XLM-R on Europarl as well as on WMT and testing on WMT.

Language Pair	HT (Europarl)	HT (WMT)	NMT (WMT)
de↔fr	0.031	0.055	0.055
cs↔en	0.116	0.336	0.325
de↔en	0.113	0.356	0.444
Macro-Avg.	0.087	0.249	0.273

Table 18: Bias values across Europarl and WMT test sets.

Model and Data

In addition to the supervised system trained on WMT16 data, we train a supervised system on a subset of Europarl (Koehn, 2005; Ustaszewski, 2019), a corpus consisting of parallel text from the proceedings of the European Parliament. We train a system for each language pair de↔fr, cs↔en, de↔en as described in Subsection 4.2. We train on 10,249 samples per direction, use a validation and test set of 1,281 further samples each per direction. Additionally, we apply the systems to the corresponding WMT test data subset as in Subsection 4.2, but with WMT22 de↔fr instead of WMT22/23 en↔ru, to test the systems' cross-domain capabilities. We experiment in the same manner with several learning rates and epochs as in Appendix C and select the best performing system for each language pair based on the validation accuracy (de↔fr: lr = dynamic, epoch = 5; cs↔en: lr = dynamic, epoch = 5; de↔en: lr = 1e-05, epoch = 5).

Results

Table 17 shows the results for the unsupervised approach, as well as both the WMT- and Europarl-based supervised systems. While the systems reach higher in-domain accuracies for en↔cs and en↔de than the WMT-based system, for de↔fr the scores are substantially lower. The results decrease further when the system is tested on out-of-domain data, namely the WMT newstest data. On average, a decrease

of 9.37% for WMT human translations and 9.18% NMT-based WMT translations is observable. These results on cross-domain applicability for supervised translation direction detection align with previous work (Sominsky and Wintner, 2019). A further notable observation, when comparing the in-domain to out-of-domain results, is the increased directional bias, which triples when the systems are applied to out-of-domain data (Table 18).

F Comparison of Results for Individual WMT Datasets

	M2M-100-418M			SMaLL-100			NLLB-200-1.3B		
Language Pair	WMT16	WMT22	WMT23	WMT16	WMT22	WMT22	WMT16	WMT22	WMT23
en↔cs	66.43	73.44	71.77	64.44	71.76	67.57	50.72	67.09	62.77
en↔de	61.88	71.65	-	63.34	71.53	-	46.53	70.27	-
en↔ru	63.91	73.34	74.05	62.29	72.28	72.83	44.30	65.20	70.60
en↔uk	-	78.98	74.38	-	77.34	73.68	-	67.53	67.85
en↔zh	-	76.18	75.60	-	75.97	75.52	-	51.97	58.87
cs↔uk	-	75.39	72.09	-	75.47	76.63	-	72.89	70.08
de↔fr	-	69.58	-	-	71.82	-	-	68.48	-
Macro-Avg.	64.08	74.08	73.58	63.36	73.74	73.25	47.19	66.20	66.03

Table 19: Average accuracy for language pairs for the individual WMT datasets and tested models.

Results by WMT year and model

The WMT16 test data predates the release of all models in our experiments. Additionally, the WMT22 data was released before the NLLB model. This circumstance raises the possibility of data contamination, as the models may have been exposed to test data during training. To address this concern, we compare results across different WMT datasets and add test samples from WMT23, which was released after all three models. If the accuracies for a dataset predating the release of a model are noticeably higher than for a more recent dataset, it would suggest data contamination. However, as shown in 19, this is not observed, indicating that our results are likely unaffected by data contamination.

G Data Statistics

Test Set	Direction	Source		Reference		
		Sents	Docs ≥ 10	HT	NMT	Pre-NMT
WMT16	cs→en	1499	40	*1499*	*1499*	16489
WMT16	de→en	1499	55	*1499*	*1499*	13491
WMT16	en→cs	1500	54	*1500*	*3000*	27000
WMT16	en→de	1500	54	*1500*	*4500*	18000
WMT16	en→ru	1500	54	*1500*	*3000*	15000
WMT16	ru→en	1498	52	*1498*	*1498*	13482
WMT22	cs→en	1448	129	2896	15928	-
WMT22	cs→uk	1930	13	1930	23160	-
WMT22	de→en	1984	121	3968	17856	-
WMT22	de→fr	1984	73	1984	11904	-
WMT22	en→cs	2037	125	4074	20370	-
WMT22	en→de	2037	125	4074	18333	-
WMT22	en→ru	2037	95	2037	22407	-
WMT22	en→uk	2037	95	2037	18333	-
WMT22	en→zh	2037	125	4074	26481	-
WMT22	fr→de	2006	71	2006	14042	-
WMT22	ru→en	2016	73	2016	20160	-
WMT22	uk→cs	2812	43	2812	33744	-
WMT22	uk→en	2018	22	2018	20180	-
WMT22	zh→en	1875	102	3750	22500	-
WMT23	cs→uk	2017	99	2017	26221	-
WMT23	en→cs	2074	79	2074	31110	-
WMT23	en→ru	2074	79	2074	24888	-
WMT23	en→uk	2074	79	2074	22814	-
WMT23	en→zh	2074	79	2074	31110	-
WMT23	ru→en	1723	63	1723	20676	-
WMT23	uk→en	1826	66	1826	20086	-
WMT23	zh→en	1976	60	1976	29640	-

Table 20: Detailed data statistics for the main experiments. Cursive: data used for validation.

Direction	Sentence Pairs
bn↔hi	1012
de↔hi	1012
cs↔uk	1012
de↔fr	1012
zh↔fr	1012
xh↔zu	1012

Table 21: Statistics for the FLORES-101 (devtest) datasets, where both sides are human translations from English.

H Sentence Length

Language Pair	0–19	20–39	40–59	60–79	80–99	100–119	120–139	140–159	160–179	180–199
NMT en-cs	59.41	62.46	65.88	69.62	74.42	75.28	74.93	78.92	79.08	81.45
NMT en-de	63.97	67.20	66.24	71.99	76.38	76.21	79.26	79.34	80.50	84.60
NMT en-ru	64.82	62.99	68.73	71.80	75.44	76.50	79.42	79.12	81.16	81.33
NMT en-uk	64.59	67.58	72.44	75.31	78.63	79.67	81.06	78.85	84.22	79.90
NMT en-zh	74.32	73.82	77.00	81.31	82.84	83.75	84.33	83.49	86.23	87.12
NMT cs-uk	61.93	69.72	76.24	76.39	80.21	80.69	83.42	81.46	83.07	84.11
NMT de-fr	57.44	65.25	70.20	74.27	76.20	73.70	79.93	72.60	76.90	78.63
Macro Avg.	63.78	67.00	70.96	74.38	77.73	77.97	80.34	79.11	81.59	82.45

Table 22: Translation direction detection accuracy by number of characters in the source sentence for different language pairs using SMaLL-100.

I Example for Forensic Dataset

	DE→EN	EN→DE
DE: Nach 30 sec. wurde das trypsinhaltige PBS abgegossen, und die Zellen kamen für eine weitere Minute in den Brutschrank.		
EN: After 30 sec, the trypsin-containing PBS was poured off, and the cells were placed in the incubator for another minute.	0.442	0.285

Table 23: Example of two segments from the forensic case (Section 5.7). M2M-100 assigns a higher probability to the English sentence conditioned on the German sentence than vice versa, suggesting that the English sentence is more likely to be a translation of the German sentence.