Adequacy or Fluency: Domain-Specific Knowledge Transfer in Multilingual Machine Translation

Stan Fris[1], Fabian Westerbeek[1], Morris de Haan[1], Julian Bibo[1], Quinten van Engelen[1],
[1]University of Amsterdam,

Abstract

Multilingual Neural Machine Translation (MNMT) models enable translation across multiple languages using a single model, yet their ability to transfer domain-specific knowledge between languages remains underexplored. This work investigates how fine-tuning transformer-style MNMT models on biomedical data affects domain-specific translation quality across languages of differing similarities. We fine-tune Qwen2-1.5B-Instruct on several domain-specific subsets of the WMT'22 Biomedical Translation corpus and evaluate performance on both a test split of WMT'22 and the term-level dataset Mantra. We propose a novel metric, Term Matching Score (TMS), to assess a model's domain knowledge by its ability to correctly translate domain-specific terminology. Our results show that domain-specific fine-tuning improves fluency-oriented metrics (ChrF++, BLEU, COMET) across most language pairs, but gains in term-level adequacy remain limited. Cross-language improvements are observed even among linguistically distant languages, indicating that domain fluency rather than language similarity drives knowledge transfer. Our code has been made public on GitHub[1].

1 Introduction

In Natural Language Processing, a large amount of research is focused on less than 1% of the languages that exist (Magueresse et al., 2020), leaving most languages understudied. For high-resource languages, there is vast amount of training data available for varied domains, however, for languages with fewer resources, the amount of training data for specific domains can be limited.

Neural models are often needed for a single specific application in which available data is also scarce, such as assistance in translating medical documents. Chu et al. (2017) have shown that fine-tuning a general model after training works well for improving domain-specific performance.

Recent advancements in the use of transformer models have allowed for the development of models which allow for translation of multiple languages with the same model, also known as Multilingual Neural Machine Translation (MNMT) (Tan et al., 2019; Ha et al., 2016).

Cooper Stickland et al. (2021) found that multilingual domain knowledge transfer in MNMT can improve results for languages absent in the source text. To the best of our knowledge, it currently is not investigated to what these changes can be attributed, as domain-specific performance is only measured using the general performance scores over the whole sequence, using metrics such as BLEU. We perform a comprehensive analysis of domain transfer for languages other than the target language in MNMT.

In evaluation of language there is an important distinction between fluency and adequacy. Adequacy measures how much of the meaning of the original sentence is preserved, whereas fluency measures how fluent (and grammatical) the translation is (Snover et al., 2009). This is a relevant consideration, as for high-impact tasks such as translation of medical documents it can be more important to have correct translations rather than fluent translation. To further investigate this, we introduce in-context term-level evaluation, which allows us to measure the efficacy of models in translating terms relevant to the task.

Additionally, in current research no specific analysis is done for the similarity of languages. Dhamecha et al. (2021) and Khemchandani et al. (2021) show that the similarity of languages is an important measure for passive knowledge transfer when fine-tuning MNMT models in cross-language domain adaptation. However, they do not provide a comprehensive analysis including complex per-

[1]https://github.com/julianbibo/multilingual-finetuning

formance metrics other than BLEU. We investigate the improvement of domain transfer when evaluating on languages from the same language family as the fine-tuning language when compared to other languages.

RQ 1: To what extent does language-specific domain adaptation in MNMT affect domain-specific translation performance for other languages in the model?

RQ 2: To what extent can cross-language transfer of domain-specific performance be attributed to improved domain knowledge?

RQ 3: How does the similarity of languages affect the transfer of domain-specific knowledge in MNMT?

2 Method

2.1 Task Definition

Machine translation considers the task of translating a sequence of tokens $S = (s_1, \ldots, s_{L_S})$ from source language $\mathcal{S}$ into a sequence of tokens $T = (t_1, \ldots, t_{L_T})$ of target language $\mathcal{T}$ (Li et al., 2023). Formally, NMT models learn the conditional probability distribution

$$P(T \mid S; \theta) = \prod_{j=1}^{L_T} P(t_j \mid t_{<j}, S; \theta).$$

Here, L_S and L_T denote the lengths of the source and target sequences, respectively, and θ are the model parameters. We can then train a model by optimizing a loss function such as the cross-entropy, which measures the negative log-likelihood of the target sequence given the source. At inference time, the target sequence is generated autoregressively until an EOS token is produced.

2.2 Models

For all experiments, we use the model Qwen2-1.5B-Instruct (Bai et al., 2023). Qwen2-1.5B-Instruct is a transformer-based, decoder-only large language model trained on diverse domains and tasks, including web documents, books, code, and encyclopedic text in over 30 languages. The architecture of Qwen is based on a LLaMA-style transformer, and the Instruct variant has been fine-tuned to follow instructions through supervised training, making it more suitable for task-oriented applications such as machine translation.

2.3 Metrics

For the evaluations of the translation models, we employ three metrics: ChrF++ (Popović, 2017), COMET (Rei et al., 2020), and BLEU (Papineni et al., 2002). These metrics provide a comprehensive view, both of surface-level overlap and semantic agreement, and to ensure both reference-based and reference-free settings are used to evaluate translation quality.

ChrF++ is a reference-based metric built upon ChrF (Popović, 2015) by combining character- and word-level n-gram F-scores. These scores are used to express agreement at both the character level and the word level, ensuring the quality is evaluated at both levels. However, since ChrF++ depends on surface overlap with the reference translation, it may undervalue correct alternative phrasing or be affected by tokenization inconsistencies in morphologically rich languages.

We also include BLEU (Papineni et al., 2002), which similarly measures n-gram overlap using modified precision and a brevity penalty. Although BLEU is less sensitive to meaning and favors surface similarity, it remains widely used in machine translation research.

To complement these surface-based metrics, we use COMET, a neural metric trained on human quality assessments. COMET predicts a scalar score based on the encodings of the source, target, and reference sentence. Its reliability largely depends on the domain of its training data. Therefore, it may misrepresent specialized terminology in biomedical contexts.

To evaluate a model's knowledge of domain-specific terminology, we introduce the Term Matching Score (TMS), which tests if the model uses correctly translated terminology in its predictions. This faithfully assesses domain-specific knowledge, ignoring features such as fluency. This metric is computed in two steps.

First, we annotate the biomedical terms in the source and reference abstracts in the evaluation dataset using DeepSeek-V3.2-Exp[2]. The prompt used for this task is provided in Appendix A. After annotation, the faithfulness of the response is verified by ensuring the existence of the generated terms in both the source and reference texts. Second, we compute the TMS. Specifically, for each document d, let T_d denote the set of annotated do-

[2] https://github.com/deepseek-ai/DeepSeek-V3.2-Exp/blob/main/DeepSeek_V3_2.pdf

main terms and $\hat{y}_d$ the model's predicted translation. A term is considered correctly translated if its exact surface form appears in $\hat{y}_d$. Let h_d be the number of correctly translated terms and $t_d = |T_d|$ the total number of annotated terms in document d. The micro-averaged TMS is then defined as the fraction $\sum_d h_d / \sum_d t_d$, reflecting the model's overall ability to preserve domain-specific terminology.

2.4 Datasets

We aim to evaluate translation performance within the domain-specific context of biomedical translation. To this end, we employ the WMT'22 Biomedical Translation corpus (Neves et al., 2022), which comprises abstracts of scientific papers sourced from the MEDLINE[3] database with parallel translations between English and French, German, Italian, and Russian. Sequence lengths vary across language pairs, reaching up to 1,449 words (English–Spanish) and 825 words (English–Russian). To standardize dataset sizes, we apply a 99th-percentile cut-off, corresponding to 400 words.

Preliminary experiments showed that fine-tuning solely on domain-specific data causes catastrophic forgetting for unseen languages (Saunders and DeNeefe, 2024). To mitigate this, we include general-domain data from the WMT'24++ corpus (Deutsch et al., 2025), which covers literary, news, social media, and speech domains. To balance fine-tuning both domain-specific and general-domain datasets are restricted to 960 samples per pair.

Evaluation uses 128 randomly selected test samples per language pair from the domain-specific data. To assess medical term translation, we use the Mantra annotated corpus (Kors et al., 2015), converting its labelled sentence pairs into a term-level dataset of source–target medical terms.

To evaluate term-specific performance on medical data, we used the Mantra annotated corpus (Kors et al., 2015), which provides source–target sentence pairs with labelled medical terms for translation. These annotations were transformed into a term-level dataset, where each medical term is paired with its corresponding translated term.

[3] https://www.nlm.nih.gov/medline/medline_overview.html

3 Experimental Setup

Our experiments were run using Python 3.9.18 and PyTorch on the Snellius HPC cluster[4].

3.1 Hyperparameters

All fine-tuning is done using the Hugging Face Trainer module[5]. Except for the learning rate, all parameters are kept at their default values, including the optimizer (AdamW) and weight decay (0.01).

The learning rate for the Trainer was determined empirically. We found that a value of $\alpha_T = 10^{-3}$ yielded the highest ChrF++ score across validation experiments.

For experiments involving LoRA fine-tuning, we follow the configuration guidelines from the Hugging Face LoraConfig documentation[6] and the associated course material[7], with minor adjustments. Specifically, we set both the LoRA rank and the scaling factor to 4, and the dropout rate to 0.1.

For prompting the language models, we adopt the translation prompt proposed by Hendy et al. (2023): *"Translate this from {lang from} to {lang to}: {lang from}: {source text} {lang to}: "*

4 Results

4.1 RQ1: Cross-language Knowledge Transfer

For this experiment, the model was fine-tuned separately for each source–target language pair using the datasets described in Section 2.4, which combine domain-specific data for the fine-tuned language pair with general-domain data for all remaining language pairs. A baseline experiment was performed using only general data. After fine-tuning, the model was re-evaluated across language pairs. This allows us to assess whether domain-specific knowledge from one language pair transfers to others within the multilingual model. The experiments were performed using english as target language, experiments with english as source language can be found in Appendix C, these results also support the observations and conclusions made in this section.

[4] Snellius hardware specifications can be found at https://servicedesk.surf.nl/wiki/spaces/WIKI/pages/30660208/Snellius+hardware.

[5] Documentation: https://huggingface.co/docs/transformers/main_classes/trainer

[6] https://huggingface.co/docs/peft/main/en/package_reference/lora#peft.LoraConfig

[7] https://huggingface.co/learn/llm-course/en/chapter11/4#lora-configuration

Table 1: Cross-language knowledge transfer evaluation for $X \rightarrow$ EN. Metric scores are shown for models fine-tuned on general data only (Baseline) and models fine-tuned on general plus language-specific biomedical data.

Finetuning Data	Metric			
FT Type	**ChrF++**	**Comet**	**BLEU**	**TMS**
Evaluated on DE → EN				
Baseline	46.21	0.791	14.36	0.110
FT on DE → EN	49.20	0.788	**17.04**	0.105
FT on FR → EN	**50.17**	0.792	16.81	0.113
FT on IT → EN	47.43	**0.798**	13.81	0.098
FT on RU → EN	49.36	0.795	16.03	**0.114**
Evaluated on FR → EN				
Baseline	55.11	0.825	22.09	0.078
FT on DE → EN	56.09	0.820	23.81	0.074
FT on FR → EN	57.80	0.823	24.61	**0.088**
FT on IT → EN	54.18	0.819	21.71	0.068
FT on RU → EN	**58.33**	**0.835**	**26.04**	0.083
Evaluated on IT → EN				
Baseline	48.69	0.795	16.80	0.103
FT on DE → EN	**53.01**	0.792	**23.58**	0.096
FT on FR → EN	52.01	0.792	18.57	**0.114**
FT on IT → EN	46.48	0.781	14.15	0.093
FT on RU → EN	51.34	**0.797**	18.47	0.105
Evaluated on RU → EN				
Baseline	49.62	0.779	15.63	0.031
FT on DE → EN	51.90	**0.784**	**17.09**	**0.033**
FT on FR → EN	**52.57**	0.783	**17.09**	0.031
FT on IT → EN	48.77	0.782	14.23	0.021
FT on RU → EN	50.90	**0.784**	16.64	0.024

Table 1 shows the results for the first experiment, where the target language was set to English. For the metrics ChrF++, COMET and BLEU we find that fine-tuning on biomedical data leads to improved performance for the source language for all languages except Italian. Furthermore, we find that results for other languages are also generally improved, for all languages except Italian, BLEU and ChrF++ scores are improved when fine-tuning with biomedical on another language compared to only fine-tuning with general data. We can therefore determine that there is a positive effect of knowledge transfer with domain-specific fine-tuning in MNMT for this setting.

4.2 RQ2: Attribution to Domain Knowledge

For this experiment we use the TMS metric and the Mantra dataset, allowing us to specifically evaluate the correct translation of biomedical terms. While ChrF++ and COMET provide reliable estimates of overall translation quality, they do not directly capture terminology accuracy. The performance of the model using the TMS was measured for the baseline models and those fine-tuned on domain-specific data, providing insight into whether performance improvement can be attributed to improved domain knowledge rather than improvements in fluency. We outline the correlation between the values of the metrics with the TMS in Appendix B.

The results for the TMS are presented in Table 1. Experiments with English as the source language suggest that TMS do not significantly improve with domain-specific fine-tuning. In evaluation on the German–English dataset, TMS remain largely unchanged across fine-tuned models, except for the Italian–English model, which shows a decrease in TMS. A similar pattern appears in the French–English dataset, however, the models fine-tuned on French–English and Russian–English exhibit a slight improvement. Evaluations on the Italian–English and Russian–English datasets follow this same trend, with some models performing marginally better than others. Overall, domain-specific fine-tuning does not appear to have a significant impact on EM performance.

Experiments on the Mantra dataset, presented in Table 2, show a drop in both ChrF++ scores for models fine-tuned with German as the target language when evaluated on French-English, Italian-English, and German–English. In contrast, these models show an improvement for the model fine-tuned on the French–English dataset, except for the German fine-tuned model, which again decreases in ChrF++. The Russian-English dataset behaves differently, showing an increase in Chrf++ when evaluated on both fine-tuned models.

4.3 RQ3: Effect of Familiarity

To investigate familiarity, we grouped the used languages according to their degree of similarity. In our language set, French and Italian are most closely related with both being a member of the Romance group. German, being part of the Germanic branch, is more distantly related to those two, yet still shares some similarities. Russian, on the other hand, belongs to the Slavic branch and is least similar to the others (Heggarty et al., 2023; Heeringa et al., 2023). We then examined whether languages that are more closely related show greater improvements in domain-specific translation performance compared to more distant languages.

The performance of fine-tuned models relative to the baselines presented in Table 2 suggests that

Table 2: Relative performance compared to baseline on WMT'22 and Mantra for TMS and ChrF++ with English as the target language. The columns indicate the source language on which the model was fine-tuned (FT) and the rows the source language that was evaluated (Ev).

Ev \ FT	Fr	It	De	Ru
WMT'22 ΔTMS%				
Fr	+1.0%	-1.0%	-0.4%	+0.5%
It	+1.1%	-1.0%	-0.7%	+0.2%
De	+0.3%	-0.2%	-0.5%	+0.4%
Ru	± 0.0%	-1.0%	+0.2%	-0.7%
Mantra ΔChrF++%				
Fr	+3.8%	+10.8%	-15.4%	+3.3%
De	-6.1%	-2.9%	-21.9%	+9.0%

the domain-specific knowledge transfer may not be correlated with language similarity. For example, fine-tuning on Russian leads to metric improvements across all languages, except Russian. Additionally, fine-tuning on German generally leads to a decrease in performance, except for Russian, even though Russian is the least related language.

What we observe instead, is that fine-tuning on a specific language pair is either consistently beneficial or hurtful to most languages. For example, fine-tuning on French or Russian results in mostly substantial performance improvements, whereas fine-tuning on Italian or German mostly in significant performance degradations.

Lastly, the results show higher variance on the Mantra dataset compared to the WMT'22 dataset. Specifically, performance changes on Mantra range from -21.9% to $+10.8\%$, whereas on WMT'22 they range only from -1.0% to $+1.1\%$. This discrepancy is likely because the models were fine-tuned on the WMT'22 training data and never exposed to samples from Mantra. Consequently, the models are less familiar with the distribution underlying the Mantra dataset, leading to greater variability in performance.

5 Discussion

The results of multilingual, domain-specific fine-tuning indicate that performance varies across language pairs. Italian generally performs on par with or worse than the baseline, except when fine-tuned in French, a closely related Romance language, suggesting that linguistic relatedness may offer some benefit. Interestingly, improvements were also observed for unrelated languages such as Russian, implying that domain knowledge itself may play a bigger role than language similarity in cross-lingual transfer. This suggests that the model likely becomes more fluent in domain-specific language usage, rather than more knowledgeable in domain-specific knowledge itself.

Furthermore, the TMS, used to assess whether domain transfer can be attributed to domain-specific knowledge, was consistently low for both fine-tuned models and baselines. The strict matching condition and the reliance on LLM-based annotation may contribute to the generally low scores. Moreover, because fine-tuning did not lead to significant changes in exact match performance, adequacy may be less affected by domain adaptation than fluency, or alternatively, the exact term match metric may be too strict and therefore less indicative of domain knowledge transfer.

The TMS results of the Mantra dataset show that performance varies substantially, performance fluctuates as a result of domain adaptation. This further supports the notion that fine-tuning likely improves model fluency in medical text more than its adequacy.

Another remark is that COMET and BLEU scores are higher when translating French-English. This may suggest that the Qwen model is slightly biased towards French medical data compared to other languages in the model.

In conclusion, this study explored multilingual domain adaptation in NMT, focusing on how fine-tuning affects performance across languages and domains. The findings suggest that while linguistic relatedness can offer some benefit, particularly between closely related languages such as French and Italian, domain knowledge itself appears to play a larger role in cross-lingual transfer. While improvements in general translation metrics such as ChrF++, BLEU, and COMET indicate enhanced fluency in domain-specific contexts, low and inconsistent term matching scores suggest that adequacy remains a challenge for the model.

Overall, our results suggest that current fine-tuning methods tend to improve fluency and style rather than precise cross-language domain knowledge transfer. Future work should explore more nuanced evaluation methods, and fine-tuning strategies aimed at improving the adequacy of NMT models in domain-specific multilingual scenarios.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen Technical Report. *arXiv*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Asa Cooper Stickland, Alexandre Berard, and Vassilina Nikoulina. 2021. Multilingual domain adaptation for NMT: Decoupling language and domain information with adapters. In *Proceedings of the Sixth Conference on Machine Translation*, pages 578–598, Online. Association for Computational Linguistics.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *Preprint*, arXiv:2502.12404.

Tejas Dhamecha, Rudra Murthy, Samarth Bharadwaj, Karthik Sankaranarayanan, and Pushpak Bhattacharyya. 2021. Role of Language Relatedness in Multilingual Fine-tuning of Language Models: A Case Study in Indo-Aryan Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8584–8595, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Wilbert Heeringa, Charlotte Gooskens, and Vincent J. van Heuven. 2023. Comparing germanic, romance and slavic: Relationships among linguistic distances. *Lingua*, 287:103512.

Paul Heggarty, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechosław Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, Henrik Liljegren, Richard F. Strand, Geoffrey Haig, Martin Macák, Ronald I. Kim, Erik Anonby, Tijmen Pronk, Oleg Belyaev, Tonya Kim Dewey-Findell, and 14 others. 2023. Language trees with sampled ancestors support a hybrid model for the origin of indo-european languages. *Science*, 381(6656):eabg0818.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv*.

Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. 2021. Exploiting language relatedness for low web-resource language model adaptation: An Indic languages study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1312–1323, Online. Association for Computational Linguistics.

Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the mantra gsc. *Journal of the American Medical Informatics Association*, 22(5):948–956.

Bryan Li, Mohammad Sadegh Rasooli, Ajay Patel, and Chris Callison-burch. 2023. Multilingual bidirectional unsupervised translation through multilingual finetuning and back-translation. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 16–31, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *ArXiv*, abs/2006.07264.

Mariana Neves, Antonio Jimeno Yepes, Amy Siu, Roland Roller, Philippe Thomas, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, Giorgio Maria Di Nunzio, Federica Vezzani, Christel Gerardin, Rachel Bawden, Darryl Johan Estrada, Salvador Lima-Lopez, Eulalia Farre-Maduel, Martin Krallinger, Cristian Grozea, and Aurelie Neveol. 2022. Findings of the wmt 2022 biomedical translation shared task: Monolingual clinical case reports. In *Proceedings of the Seventh Conference on Machine Translation*, pages 694–723, Abu Dhabi. Association for Computational Linguistics.

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

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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

Danielle Saunders and Steve DeNeefe. 2024. Domain adapted machine translation: What does catastrophic forgetting forget and why? *arXiv preprint arXiv:2412.17537*.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece. Association for Computational Linguistics.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 963–973, Hong Kong, China. Association for Computational Linguistics.

A Term annotation prompt

Below is the prompt used for the annotation of (bio)medical terms in the source and reference text:

You are a precise biomedical term aligner. Given a source and target sentence, identify biomedical/scientific, domain-specific terms (single or multiword: diseases, syndromes, procedures, drugs, genes, pathologies, anatomy) that appear in the source, and return their exact translations that appear in the target. Only include pairs if both surface forms appear verbatim in their respective sentences. Output strictly JSON with a 'pairs' array of objects: "pairs":["src":"...","tgt":"...", ...] and nothing else.

B Metric correlation

In Figure 1, we show the correlation of the TMS metric and the other metrics. We see that none of the metrics are strongly correlated with the TMS score. This supports our claim that our metric is able to provide novel information on the text presented.

C Domain-adaptation fine-tuning with English source

Our general experiments included english as a target language, as this showed better initial results. For completeness, we also conduct our main experiment using english as a source language, as seen in Table 3. Here, we observe similar conclusions compared to english target fine-tuning. Notably, the performance increase when fine-tuning on French is very high. Overall, we again observe a high-level of cross-language domain transfer.

Table 3: Cross-Language Knowledge Transfer Evaluation for EN → X (Source-Side Finetuning). Performance metrics (ChrF++, Comet, BLEU, and TMS) are shown for models finetuned on general plus language-specific biomedical data from the **Source-Side** domain (FT Lang. EN → X).

Finetuning Data	**Metric**			
FT Type	**ChrF++**	**Comet**	**BLEU**	**TMS**
Evaluation on EN → DE				
Baseline	38.908	0.667	7.557	0.202
FT on EN → DE	41.107	0.637	8.111	0.189
FT on EN → FR	**45.698**	**0.690**	10.109	0.212
FT on EN → IT	44.086	0.671	9.748	0.219
FT on EN → RU	44.567	0.682	**10.259**	**0.236**
Evaluation on EN → FR				
Baseline	47.399	0.775	15.264	0.259
FT on EN → DE	49.162	0.786	16.288	0.259
FT on EN → FR	**54.762**	**0.804**	**20.750**	**0.282**
FT on EN → IT	51.497	0.780	17.906	0.259
FT on EN → RU	50.961	0.779	17.782	0.281
Evaluation on EN → IT				
Baseline	40.857	0.713	8.638	0.182
FT on EN → DE	45.708	0.726	11.241	0.198
FT on EN → FR	**50.446**	**0.759**	**13.978**	**0.224**
FT on EN → IT	45.865	0.721	11.012	0.181
FT on EN → RU	45.436	0.721	11.956	0.208
Evaluation on EN → RU				
Baseline	41.235	0.789	9.454	0.150
FT on EN → DE	44.311	0.789	10.502	0.158
FT on EN → FR	45.594	**0.795**	**12.035**	**0.172**
FT on EN → IT	43.346	0.777	10.653	0.159
FT on EN → RU	**46.066**	0.781	11.561	0.164

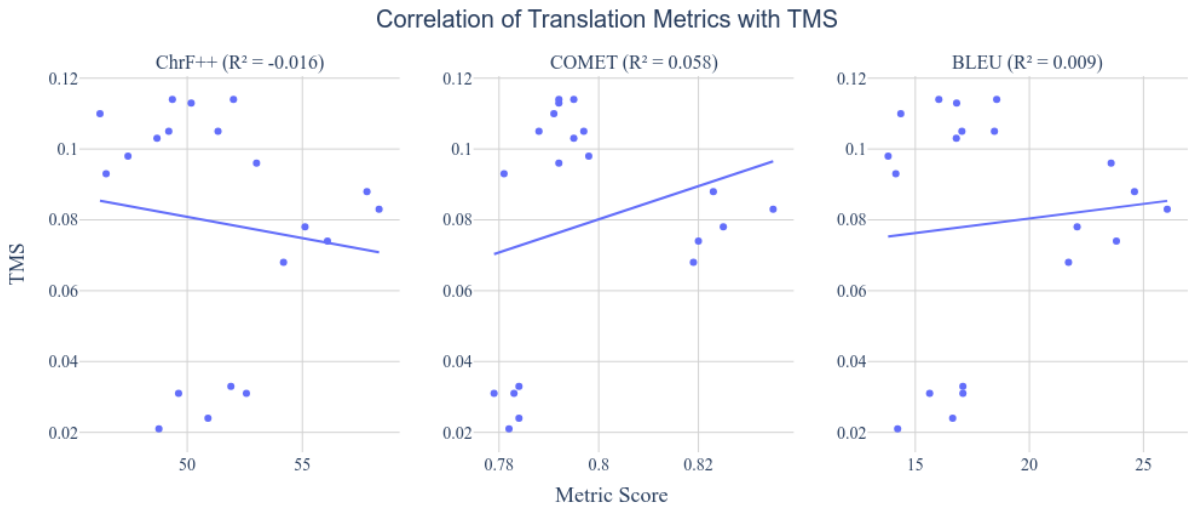

Figure 1: Correlation between TMS and several translation metrics with the Pearson correlation coefficient. All three metrics are uncorrelated to the TMS metric.