Linguistic Complexity in Domain-Specific Neural Machine Translation

Daria Sokova[1][0009−0003−5486−1527] and Cristina Toledo-Báez[1][0000−0003−0604−797X]

[1] University of Málaga, Spain
dariasokova@gmail.com
toledo@uma.es

Abstract. This study investigates the impact of linguistic complexity in domain-specific data on NMT performance. Using a data selection approach, we created a dataset and trained Transformer models both on domain-specific data and on a general dataset. Linguistic complexity analysis reveals that the domain-specific dataset exhibits notable linguistic complexity, with syntactic, lexical, and textual characteristics posing challenges for NMT. Training models on this complex data resulted in lower translation quality compared to models trained on general data, particularly evident when translating into a morphologically rich language like Russian. These findings may suggest that linguistic complexity in domain-specific data can present challenges for NMT performance, making training computationally intense and hindering the model's ability to learn accurate representations. Our study emphasizes the importance of considering data characteristics in NMT training, especially in low-resource scenarios.

Keywords: Domain-specific Neural Machine Translation · Data Complexity · Lexical Characteristics.

1 Introduction

Neural Machine Translation (NMT) has emerged as a state-of-the-art approach for automatic translation, largely supplanting statistical machine translation approaches (Stahlberg, 2020). NMT systems offer cost effectiveness and the opportunity to leverage large-scale parallel corpora and neural network architectures to train models that can learn the patterns of natural language. However, NMT systems still face a number of challenges that affect their performance. These challenges include difficulties in handling out-of-domain translation and rare words, problems with long sentences and limitations in capturing linguistic nuances and semantic information from training data, especially when the data is scarce. Moreover, the complexity and quality of training data has also been found to impact NMT training and performance (Agrawal and Singh, 2023; Arora et al., 2021).

Proceedings of New Trends in Translation and Technology, pages 191–200
3–6 July, 2024, Varna, Bulgaria.
https://doi.org/10.26615/issn.2815-4711.2024_015

In this study, we focus on the impact of linguistic characteristics of domain-specific data on NMT training and inference. We conduct experiments by creating a small domain-specific dataset through data selection and training Transformer models on general and in-domain data, analyzing the effects of linguistic complexity and domain specificity on translation quality.

2 Related Work

Although NMT has become the state-of-of-the-art approach to automatic translation, there are still a number of challenges that NMT systems face. These challenges tend to particularly manifest themselves under low-resource conditions (Koehn and Knowles, 2017; Chakrabarty, 2020).

Koehn and Knowles (2017) identify out-of-domain words, long sentences and word alignment among the challenges for NMT systems that lead to errors in the output. They report that NMT systems face difficulties when tackling out-of-domain translation, where the input or source text being translated falls outside the domain or subject matter that the NMT system was trained on. When dealing with out-of-domain input, NMT systems are prone to hallucinating and sacrificing adequacy for the sake of fluency. In addition to that, they find that when translating rare words, NMT systems tend to fall back on deletion. The authors also demonstrate that NMT systems struggle when handling highly inflected rare words and lexical units not observed or observed only once during the training, particularly named entities.

Brussel et al. (2018) also report a number of challenges for NMT systems. They find that NMT output reveals a high number of mistranslations that are not semantically related to the source, which is uncommon for phrase-based and rule-based machine translation models. Consistent with prior studies (Wu et al., 2016; Koehn and Knowles, 2017) they highlight that NMT systems provide sufficiently fluent output to seemingly mask the omissions in meaning. They also confirm findings by Koehn and Knowles (2017) in relation to the NMT performance degrading when handling proper nouns.

Focusing on the impact of overall data complexity, Agrawal and Singh (2023) examine whether enhancing the complexity rather than the size of the training corpus for language models can lead to improved performance on downstream tasks and find a significant correlation between corpus complexity, similarity to downstream data and task performance. (Arora et al., 2021) reveal that extensive noise filtering and normalization significantly enhanced machine translation performance, particularly for neural models, despite the reduction in corpus size, indicating the critical role of data quality in translation accuracy.

Following the findings of previous studies, our research is focused on identifying the complexity level and linguistic characteristics of domain-specific data obtained through data selection and its impact on NMT training and translation performance.

3 Materials and Methods

Parallel Data In order to evaluate the impact of linguistic characteristics of the training data on the performance of NMT systems, we aimed to create conditions of a low-resource domain-specific scenario. Although both Russian and English are considered high-resource languages, this language pair presents certain difficulties for NMT, particularly in handling the rich morphology of Russian and the syntactic complexity of both languages. We aimed to create challenging conditions by selecting a relatively small dataset belonging to the domain of the environment. As no open-source English-Russian parallel datasets in the environmental domain were available at the moment of carrying out the research project, we employed a data selection tool to select the domain-specific data from a larger general corpus.

For the data selection process we employed the Python Tool for Selecting Domain-Specific Data in Machine Translation (Pourmostafa Roshan Sharami et al., 2021) that allows for selecting domain-specific data from larger generic corpora. The tool's operation requires three inputs: the source and target sides of an out-of-domain parallel corpus and a monolingual domain-specific corpus. The tool uses semantic search to find generic sentences that are similar to domain-specific seed sentences by comparing the vectors and ranking them based on their cosine similarity score. As our general parallel corpus, we used the MultiUN corpus, which is a collection of translated documents from the United Nations. For the domain-specific input, we used the PANACEA Environment English monolingual corpus, which comprises documents sourced from the internet, automatically identified as English language content, and automatically categorized as pertaining to the environmental domain.

The tool was used to extract 500,000 domain-specific parallel sentences. We filtered the acquired dataset following the pre-processing pipeline proposed by DataLitMT. The filtering steps included deleting rows that contain empty cells, deleting duplicates, source-copied rows and sentences that are too long, removing HTML code, deleting rows with empty cells and shuffling rows. The filtered dataset was tokenized using SentencePiece and split into training, development and testing sets. The training dataset contained 400,000 parallel sentences, while the development and testing sets each contained 50,000 parallel sentences. For the purpose of comparing the impact of the in-domain data extracted through data selection and out-of-domain data on NMT training and performance, 500,000 parallel sentences of the MultiUN parallel corpus were extracted and pre-processed to serve as the out-of-domain dataset.

In-domain Data Complexity In order to estimate the complexity of the in-domain data, we carried out a manual analysis in order to identify the linguistic characteristics that can potentially present challenges for NMT in a low-resource scenario. To analyze the data, we randomly rearranged the rows in the full domain-specific dataset and selected the initial 100 pairs of sentences from this shuffled dataset for manual examination.

At a textual level, the lack of coherence in the data became apparent. Rather than presenting a cohesive narrative or thematic continuity, the dataset contained a collection of sentences that are not contextually connected. On the syntactic level, long and grammatically complex sentences were prevalent in the dataset. In both English and Russian, the sentences contain complex grammatical structures, which is particularly noticeable in the Russian text. The Russian-side data is characterized by long noun phrases and convoluted sentence structures where the verb is detached from the subject by a long noun phrase. The example English sentences from the dataset sample illustrating potential translation dufficulties are presented in Table 1.

Table 1. Sentence examples from the domain-specific dataset.

Difficulty	**Example Sentence**
NE and specialized vocabulary (English)	*The Bureau of CAMI had called on the Director-General to conduct two important studies, one on mechanisms and strategies for the industrialization of Africa in the new millennium, the other on rendering CAMI sustainable.*
long noun phrase	*New scientific knowledge on dose-response relationships and on critical limits will foster progress in the further development of the critical loads/levels approach and in the dynamic assessment of changes in the environment.*

The metadata provided for the MultiUN parallel corpus does not seem sufficient to conclude whether the parallel data is fully human-translated texts or if machine translation was applied to a portion of the data. The presence of raw machine translation output in the parallel dataset can introduce noise to the training data and exacerbate the linguistic challenges. It is also important to note that there are multiple punctuation errors present in the Russian side of the parallel dataset, which can also hinder the model's learning.

On the lexical level, the data contained a wide range of specialized vocabulary, including domain-specific terminology, proper nouns, document titles and abbreviations. The terminological density of the dataset may pose challenges to the NMT model when it comes to handling out-of-vocabulary words unseen in the training data (Agrawal and Singh, 2023). High lexical density can also result in a larger vocabulary, which can significantly influence both the training process and the quality of translation. Additionally, the wide range of specialized vocabulary increases the out-of-vocabulary token rate, where the model encounters words it has not seen during training. This further complicates translation, as the model may resort to replacing out-of-vocabulary tokens with generic placeholders or providing inaccurate translations.

Thus, the closer inspection of the dataset obtained through data selection revealed a number of linguistic characteristics that can potentially impact the training process and translation performance. In order to computationally assess the data complexity, we applied a number of data complexity metrics following

Agrawal and Singh (2023). We assessed the complexity of the domain-specific and out-of-domain datasets with a number of readability metrics, including Type-token ratio (TTR), Corrected type-token ratio (CTTR), Mean segmental type-token ratio (MSTTR), Moving average type-token ratio (MATTR), Measure of Textual Lexical Diversity (MTLD) and Hypergeometric Distribution Diversity (HDD).

The TTR calculates the proportion of unique words to total words while CTTR adjusts for text length, providing more accurate diversity measures for longer texts. The MSTTR computes diversity across text segments, revealing variations within the dataset while the MATTR smooths out fluctuations in diversity over the text. The MTLD considers word sequence length and the HDD measure evaluates diversity based on the probability of encountering new words, factoring in the distribution of unique words within the text. Overall, these metrics provide insights into the lexical diversity and complexity of the data by measuring factors such as unique word proportion, text length adjustment, diversity across segments, word sequence length, and the probability of encountering new words.

According to the lexical richness scores, both datasets exhibit a substantial number of unique words, indicating a wide range of vocabulary in each. Additionally, the Russian text demonstrates a notably higher TTR compared to the English text, especially for the in-domain dataset. Similarly, the CCTR reinforces the observation of higher lexical diversity in the Russian-side data, suggesting a more varied vocabulary and potentially more complex linguistic structures. Moreover, the MTLD for the Russian data is substantially higher than such of the English text, further indicating greater lexical richness and diversity in the Russian dataset. Overall, the lexical richness metrics mainly point to overall high level of vocabulary diversity of the out-of-domain and in-domain datasets. However, the Russian data, particularly the in-domain dataset, demonstrate higher lexical diversity across multiple metrics. The results of this analysis are presented in Table 2.

Table 2. Data Complexity of the Domain-specific and Out-of-domain Datasets.

	TTR	CTTR	MSTTR	MATTR	MTLD	HDD
English ID	0.009	25.70	0.87	0.87	95.40	0.85
Russian ID	0.02	53.65	0.94	0.94	542,68	0.83
English OOD	0.007	16.85	0.85	0.85	56.51	0.83
Russian OOD	0.016	38.47	0.92	0.92	159.92	0.93

Following Agrawal and Singh (2023) in assessing the readability of textual content across languages, we used the Flesch-Kincaid Grade Level and Flesch Reading Ease metrics (Flesch, 1948) to measure the readability of the domain-specific and out-of-domain data in order to get a computational insight on yet another property of the data. For the English side of the in-domain data sample, the Flesch-Kincaid Grade Level was calculated to be approximately 14.94, indicat-

ing that the material is intelligible to individuals at a high school grade level. However, the Flesch Reading Ease score of approximately 33.69 suggests that the text falls within the category of "difficult," posing potential challenges for readers. Conversely, the assessment of the Russian side revealed a higher level of linguistic complexity. With a Flesch-Kincaid Grade Level of approximately 17.66, the content demands a more advanced level of education for comprehension. Furthermore, the Flesch Reading Ease score of approximately 3.71 underscores the text's complex nature, indicating a considerable degree of difficulty for readers. The Flesch-Kincaid Grade Level scores for both sides of the out-of-domain data sample are higher than those of the in-domain sample, potentially pointing to higher complexity. The Flesch Reading Ease scores for the English side of the out-of-domain data sample is lower than for the in-domain sample while for Russian it is marginally higher. Nevertheless, for both samples, the Flesch Reading Scores computed for Russian are rather low, indicating a low level of readability.

Table 3. Readability of the Domain-specific and Out-of-domain Datasets.

	En ID	Ru ID	En OOD	Ru OOD
Flesch-Kincaid Grade Level	14.94	17.66	18	19.38
Flesch Reading Ease	33.69	3.70	21.57	4.8

Overall, the linguistic characteristics identified in the datasets and the results of data complexity analysis point to a high level of complexity, particularly of the Russian-side data, which can potentially pose challenges for NMT performance, especially in low-resource conditions.

4 Experiment

Training a Transformer with Domain-specific Data In order to study the impact of data characteristics on the training and NMT performance quality, we trained Transformer models (Vaswani et al., 2017) using the OpenNMT-py toolkit. For these experiments, the configuration proposed by the DataLitMT tutorial was used. Vocabulary sizes for both source and target languages are set to 50,000. Source and target sequence lengths are filtered to a maximum of 150 tokens each. Early stopping is implemented to halt training if no improvement is observed after 10 validations.

Both the encoder and decoder consist of 6 layers. Each layer incorporates multi-head self-attention mechanisms with 8 attention heads. Each hidden layer within the encoder and decoder components of the Transformer has 512 units. For word embeddings, both the source-side and target-side sizes were set to 512. Training samples are batched based on token count, with a batch size of 4096 tokens and a bucket size of 262144 tokens. Validation batches consist of 2048 tokens. Gradient accumulation is performed every 4 steps, with a maximum of

2 gradient accumulation steps. Noam decay is applied for learning rate scheduling. Label smoothing with a coefficient of 0.1 is employed, and parameters are initialized with a Glorot distribution. The Adam optimizer with a learning rate of 1 is used.

We trained two Transformer models using the OpenNMT-py toolkit on the domain-specific dataset obtained with the Data Selection Tool. For both translation directions, Russian-to-English and English-to-Russian, the models were trained for 100,000 training steps. Notably, the vocabulary for the Russian-side domain-specific data was very large comprising over 240,000 tokens as opposed to the English vocabulary of around 50,000 tokens, which may be attributed to the high level of morphological complexity of the Russian language. To further investigate the influence of domain-specific data on model training, we trained two more models with the same training configuration using the out-of-domain dataset extracted from the MultiUN corpus. Notably, models trained on general-domain data terminated training before reaching 100,000 training steps, with the best-performing models occurring at step 20,000.

We evaluated the models' performance using the MATEO evaluation tool (Vanroy et al., 2023) that incorporates traditional and neural evaluation metrics. The evaluation results are presented in Table 4.

Table 4. Evaluation Metrics for Transformer Models

	bertscore	**bleurt**	**comet**	**BLEU**	**chrF2**	**TER**
EnRu_ENV	83.4	55.2	76.3	30.5	52.4	64.2
EnRu_GEN	87.0	72.1	85.5	41.3	64.3	55.3
RuEn_ENV	87.8	69.2	81.3	39.9	61.9	54.5
RuEn_GEN	88.2	72.1	82.2	42.9	64.4	51.1

It can be observed from the evaluation results that the worst-performing model is the English-to-Russian model trained on the domain-specific dataset. The Russian side of the domain-specific dataset was found to be particularly linguistically complex. In addition to that, translating into a morphologically richer language might be especially challenging for an NMT system (García-Martínez et al., 2020). Overall, the training results may indicate that more complex and lexically diverse training data can make NMT training more computationally intense as well as hinder model's ability to learn accurate representations, and, therefore, decrease translation quality.

Future Experiments The experiments with the baseline model may signify that the in-domain dataset is more challenging for the model to learn quality representations from, especially translating from English into more morphologically complex Russian. This can also be observed from the discrepancy of the vocabulary size built for the Russian and English source data. Thus, being provided with limited training data marked by challenging characteristics, the model

demonstrates worse performance. One solution for this issue may be increasing the amount of training data. However, in low-resource scenarios it is often not an option. Alternatively, addressing the challenges posed by low-resource scenarios and morphologically complex languages in NMT could involve exploring domain adaptation techniques, such as fine-tuning and mix fine-tuning pre-trained models using domain-specific data.

5 Conclusion

This study explored how linguistic complexity in domain-specific data may affect NMT training and performance. Experiments with Transformer models revealed that the NMT systems faced with overly complex and low-resource data demonstrate poorer performance. These findings highlight the need to consider data characteristics during NMT training, particularly in low-resource settings. Future solutions may involve integrating linguistic features into NMT models to address these challenges.

Acknowledgments. The authors of this paper would like to express their sincere gratitude to Dr. Constantin Orasan for his invaluable contributions to this research. This work would not be possible without his expertise, guidance, and support. The authors also express their gratitude to the anonymous reviewers for their valuable feedback and suggestions.

C. Toledo-Báez acknowledges that this work was carried out in the framework of the following research projects:

- GAMETRAPP (ref. no. TED2021-129789B-I00/AEI/10.13039/501100011033/Unión Europea NextGenerationEU/PRTR)
- NEUROTRAD (B1-2020_07)
- TRADUTEACH (PIE22-124)
- VIP II (ref. no.PID2020-112818GB-100/AEI/10.13039/501100011033)
- RECOVER (ref. no. ProyExcel_00540)
- Proof of Concept (PDC2021-121220-I00)
- T2T (D5-2023_14)

References

1. A. Agrawal and S. Singh, "Corpus Complexity Matters in Pretraining Language Models," in Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), N. Sadat Moosavi, I. Gurevych, Y. Hou, G. Kim, Y. J. Kim, T. Schuster, and A. Agrawal, Eds., Toronto, Canada (Hybrid): Association for Computational Linguistics, Jul. 2023, pp. 257–263. doi: 10.18653/v1/2023.sustainlp-1.20.
2. K. K. Arora, G. S. Tomar, and S. S. Agrawal, "Studying the Role of Data Quality on Statistical and Neural Machine Translation," in 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India: IEEE, Jun. 2021, pp. 199–204. doi: 10.1109/CSNT51715.2021.9509604.

3. A. Chakrabarty, R. Dabre, C. Ding, M. Utiyama, and E. Sumita, "Improving Low-Resource NMT through Relevance Based Linguistic Features Incorporation," in Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 4263–4274. doi: 10.18653/v1/2020.coling-main.376.
4. C. Chu and R. Wang, "A Survey of Domain Adaptation for Neural Machine Translation." arXiv, Jun. 01, 2018. Accessed: May 07, 2024. [Online]. Available: http://arxiv.org/abs/1806.00258
5. R. Flesch, "A new readability yardstick," Journal of Applied Psychology, vol. 32, no. 3, pp. 221–233, 1948, doi: 10.1037/h0057532.
6. Z. Z. Hlaing, Y. K. Thu, T. Supnithi, and P. Netisopakul, "Improving neural machine translation with POS-tag features for low-resource language pairs," Heliyon, vol. 8, no. 8, p. e10375, Aug. 2022, doi: 10.1016/j.heliyon.2022.e10375.
7. P. Koehn and R. Knowles, "Six Challenges for Neural Machine Translation," in Proceedings of the First Workshop on Neural Machine Translation, Vancouver: Association for Computational Linguistics, Aug. 2017, pp. 28–39. doi: 10.18653/v1/W17-3204.
8. Y. Pan, X. Li, Y. Yang, and R. Dong, "Multi-Source Neural Model for Machine Translation of Agglutinative Language," Future Internet, vol. 12, no. 6, Art. no. 6, Jun. 2020, doi: 10.3390/fi12060096.
9. J. Pourmostafa Roshan Sharami, D. Shterionov, and P. Spronck, "Selecting Parallel In-domain Sentences for Neural Machine Translation Using Monolingual Texts," Computational Linguistics in the Netherlands Journal, vol. 11, pp. 213–230, Dec. 2021, doi: 10.26116/5eav-qz46.
10. A. Raganato and J. Tiedemann, "An Analysis of Encoder Representations in Transformer-Based Machine Translation," in Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 287–297. doi: 10.18653/v1/W18-5431.
11. R. Sennrich and B. Haddow, "Linguistic Input Features Improve Neural Machine Translation," in Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers, Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 83–91. doi: 10.18653/v1/W16-2209.
12. F. Stahlberg, "Neural Machine Translation: A Review," Journal of Artificial Intelligence Research, vol. 69, pp. 343–418, Oct. 2020, doi: 10.1613/jair.1.12007.
13. Vanroy, B., Tezcan, A., Macken, L. (2023). MATEO: MAchine Translation Evaluation Online. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, ... H. Moniz (Eds.), Proceedings of the 24th Annual Conference of the European Association for Machine Translation (pp. 499–500). Tampere, Finland: European Association for Machine Translation (EAMT).
14. L. Van Brussel, A. Tezcan, and L. Macken, "A fine-grained error analysis of NMT, PBMT and RBMT output for English-to-Dutch," in Proceedings of the Eleventh International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), 2018, pp. 3799–3804. Accessed: May 22, 2023. [Online]. Available: http://hdl.handle.net/1854/LU-8561558
15. A. Vaswani et al., "Attention Is All You Need." arXiv, Dec. 05, 2017. Accessed: May 02, 2023. [Online]. Available: http://arxiv.org/abs/1706.03762
16. "Learning Resources - DataLitMT." Accessed: May 09, 2024. [Online]. Available: https://itmk.github.io/The-DataLitMT-Project/resources/training-an-nmt-model

17. Y. Wu et al., “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,” Sep. 2016.
18. “MultiUN.” Accessed: Jan. 23, 2024. [Online]. Available: https://opus.nlpl.eu/MultiUN.php
19. “PANACEA Environment English monolingual corpus – ELRA Catalogue.” Accessed: Jan. 24, 2024. [Online]. Available: https://catalogue.elra.info/en-us/repository/browse/ELRA-W0063/