What's Lost in Translation? Tackling Errors in Multilingual NLI Through Sequential Training

Anonymous NLP Final Project submission

Abstract

This work presents a comprehensive analysis of natural language inference (NLI) models across multilingual settings, with a particular focus on adversarial and linguistic error testing. Using curated datasets in English and Russian, we evaluate how well state-of-the-art models, such as multilingual BERT and XNLI, handle complex linguistic challenges, such as metaphor detection, negation, and high word overlap. We introduce an approach to fine-tuning that integrates adversarial examples and targeted linguistic challenges, improving model robustness across test cases. Error analysis highlights key challenges, such as difficulties with numerical reasoning and semantic similarities, offering insights into the limitations of current NLI systems. Our findings underline the importance of linguistically diverse datasets for evaluating and enhancing multilingual NLI performance.

1 Introduction

In Natural Language Inference (NLI) a model is given a sentence pair and then tasked with determining the logical relationship between the two. The model must decide whether one sentence entails the other, contradicts it, or is neutral. This task is especially challenging in multilingual settings, where models must generalize across varying linguistic structures. The XNLI dataset (Conneau et al. 2018) provides a key benchmark for evaluating cross-lingual sentence comprehension, extending the MultiNLI corpus to 15 languages. XNLI enables evaluation of models' multilingual understanding, with baselines including machine translation and multilingual encoders trained on parallel data.

While XNLI has progressed the study of multilingual NLI, its relatively simpler examples often fail to fully test models' robustness to adversarial or complex linguistic problems. In contrast, datasets like ANLI (Adversarial NLI) introduce harder examples designed through a human-and-model-in-the-loop process, that can identify vulnerabilities in even state-of-the-art models like BERT. In this work, we build on the XNLI framework to fine-tune multilingual BERT (mBERT) and evaluate its robustness in natural language understanding. Initial training and testing on XNLI achieves high performance with some discrepancies across the two languages studied (Russian and English). These reveal linguistic challenges such as metaphor misinterpretation and paraphrasing. When tested on ANLI, performance decreases significantly, indicating a lack of model generalization with out of distribution examples.

The choice of English and Russian for our experiments reflects the need to assess NLI models across linguistically and grammatically diverse environments. English is likely the most studied language in the space of natural language understanding while Russian is likely less studied and uses grammatically different syntactic structures. By including datasets targeting challenges such as metaphor and negation, we aim to expose weaknesses and limitations in model generalization.

To identify challenges, we conduct a comprehensive error analysis, categorizing model misclassifications into specific error types such as numerical reasoning and negation misunderstanding. Additionally, we explore targeted fine-tuning strategies, including the use of adversarially generated data to mitigate specific weaknesses. By adding to the large corpus of work started by XNLI and extending its evaluation to adversarial datasets, we hope to advance the development of robust and linguistically generalizable multilingual NLI models.

2 Method

Multilingual BERT (mBERT) is a variant of BERT (Bidirectional Encoder Representations from Transformers) designed to handle 104 languages without requiring explicit alignment between them. BERT itself is a transformer-based model that leverages bidirectional attention to capture contextual information from both directions in a sentence, achieving state-of-the-art results on a wide range of natural language processing tasks (Devlin et al. 2019). mBERT inherits this architecture, using a shared subword vocabulary to enable multilingual processing and cross-lingual transfer. Its effectiveness lies in its ability to learn generalized sentence representations that work across languages, even in zero-shot scenarios.

2.1 Training on XNLI

We fine-tuned the mBERT model on the XNLI dataset, a multilingual extension of the MultiNLI corpus (Conneau et al. 2018), over five epochs. Training performance was monitored using F1-score, precision, and accuracy on the validation set. The highest overall metrics were achieved at epoch 4, with an F1-score of 0.757, precision of 0.770, and accuracy of 0.756.

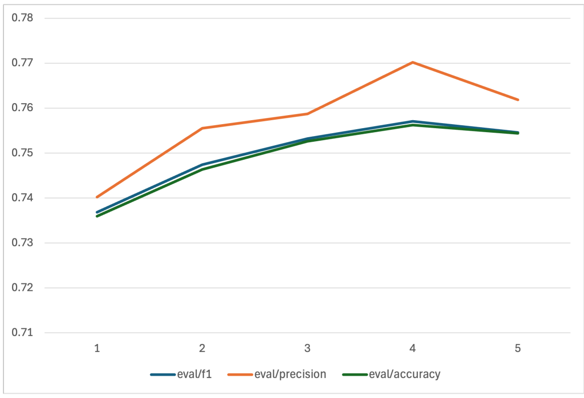
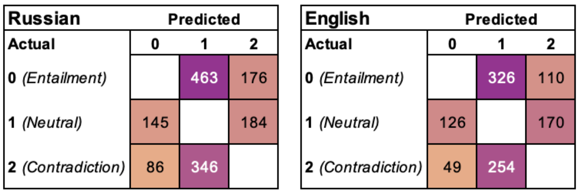


Figure 1: Validation Metrics by Epoch

The most substantial improvement in performance occurred during the first epoch, with diminishing gains in subsequent epochs. By epoch 5, the F1-score began to decline, indicating that the model was starting to overfit to the XNLI training set. The remainder of this analysis will be using the mBERT model trained over 4 epochs of XNLI.

2.2 Testing on XNLI

We evaluated the model on the XNLI test set, achieving an overall F1-score of 75.8. Performance varied across the two languages, with English achieving a higher F1-score of 79.5 compared to 72.1 for Russian. Analysis of confusion matrices reveals that the largest source of errors in Russian arose from misclassifications of true entailment and contradiction instances as neutral.



Russian	**Predicted**		
Actual	**0**	**1**	**2**
0 *(Entailment)*		463	176
1 *(Neutral)*	145		184
2 *(Contradiction)*	86	346	

English	**Predicted**		
Actual	**0**	**1**	**2**
0 *(Entailment)*		326	110
1 *(Neutral)*	126		170
2 *(Contradiction)*	49	254	

Figure 2: Confusion Matrix of Misclassifications

Further investigation uncovered that some of these errors stemmed from translation inconsistencies between English and Russian. One notable example involved a mismatch in verb gender, where the premise used a feminine verb and the hypothesis used a masculine verb, leading the model to incorrectly label the pair as neutral:

- Russian Premise: Сегодняшний вопрос напоминает мне единственный раз, когда я **ходила** в Radio City Music Hall на Рождественский танец (ходила – went (khodila, feminine singular, past tense, imperfective verb))
- Russian Hypothesis: Я **пошел** на рождественский праздник, чтобы увидеть живое Рождество (пошел – went (poshyol, masculine singular, past tense, perfective verb))
- Russian Model Prediction: Neutral
- English Premise: Today's question reminds me of the one and only time I **went** to the Radio City Music Hall Christmas Pageant for the Living Nativity. (ungendered verb)

- English Hypothesis: I **went** to the Christmas Pageant to see the living nativity. (ungendered verb)
- Actual Label: Entailment

While translation issues contributed to a small portion of the errors, a more significant challenge was the model's limited ability to understand metaphors. Although this issue was observed in both languages, it was slightly more pronounced across the Russian errors. This aligns with one of the XNLI limitations noted by Conneau et al., where cultural nuances and metaphorical expressions are not well-captured in the XNLI dataset due to its reliance on direct translation. These findings suggest that testing cross-lingual NLI models on metaphorical data presents an intriguing and challenging direction for further research.

2.3 Metaphor Testing

To evaluate our cross-lingual language model on challenging metaphorical data, we constructed a new dataset derived from the Language Computer Corporation (LCC) annotated metaphor datasets (Mohler et al. 2016). The LCC dataset, annotated over three years in four languages (English, Spanish, Russian, and Farsi), assigns a metaphoricity score to each example sentence, with a high score of 3 indicating a clear metaphor.

We began by extracting all Russian and English examples with a metaphoricity score of 3, removing duplicates arising from overlapping topics. This resulted in 1,203 Russian examples and 10,903 English examples. To ensure balanced representation across languages, we randomly sampled 1,200 English examples to match the size of the Russian dataset. The final dataset, consisting of 2,403 examples, was evenly distributed across languages.

Each example was labeled as entailment, neutral, or contradiction. To generate hypotheses, we used GPT-4 to create sentence pairs. Premises were taken directly from the LCC dataset, and hypotheses were generated based on the assigned label. For the neutral category, we introduced variety by dividing the examples into two subgroups: one prompted GPT-4 to produce completely unrelated sentences, while the other generated neutral yet contextually related hypotheses.

To ensure the quality and reliability of the dataset, we implemented a manual validation step. A random subset (20 per language) of the generated examples were reviewed by bilingual team members proficient in English and Russian. No direct errors were found in this quality check, but after this check we created two separate neutral approaches as the generated hypotheses were too similar to the premises.

The resulting dataset was split into training (0.6), validation (0.2), and test (0.2) sets. Testing the metaphor dataset against the XNLI-trained model revealed a significant drop in F1 scores, from 79.5 to 61.3 for English and from 72.1 to 49.7 for Russian. This drop highlights the challenge posed by metaphors, particularly for Russian, demonstrating the need for further adaptation to handle these linguistic phenomena effectively.

2.4 Adversarial Testing

To further assess the model's robustness, we evaluated it on the Adversarial Natural Language Inference (ANLI) dataset (Nie et al. 2020), a benchmark designed to challenge state-of-the-art language models. For multilingual evaluation, we translated all English premise-hypothesis pairs in ANLI into Russian using GPT-4 with a translation prompt.

When tested on the translated ANLI test set, the model's performance dropped significantly, achieving an overall F1-score of 29.8 (29.3 for English and 30.3 for Russian). This result underscores the dataset's complexity and its divergence from the XNLI data training distribution. To gain deeper insights, we analyzed 100 misclassifications, evenly split between English and Russian, and identified nine key error types:

- **Assumes external knowledge**: Model unable to draw sufficient conclusions as hypothesis relies on external facts not explicitly provided in the premise.
- **Overgeneralization error**: Model infers conclusions that overly broaden the scope of information presented in the premise or hypothesis.
- **Temporal ambiguity**: Model fails to resolve unclear or implied time-related re-

lationships between the premise and hypothesis.

- **Attribution error**: Model confuses or misattributes roles, actions, or characteristics described in the premise.
- **Negation misunderstanding**: Model fails to correctly interpret negations in the hypothesis and overclassifies such examples as contradictions.
- **High word overlap**: Model overemphasizes lexical similarity between the premise and hypothesis, leading to incorrect classification.
- **Confused by extra context**: Model is misled by irrelevant or distracting details in the premise.
- **Hard math or numbers**: Model misinterprets or mishandles numerical reasoning or quantitative information in the premise or hypothesis.
- **Semantic approximation**: Model fails to correctly interpret semantically similar expressions in the premise or hypothesis.

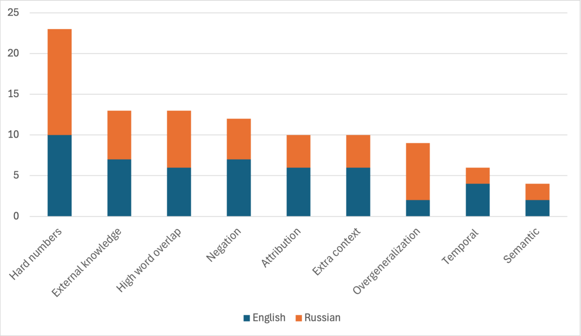


Figure 3: 100 sampled errors by type

Among these error types, the metaphor dataset already addresses some aspects of semantic approximation. Thus, subsequent fine-tuning efforts focus on negation misunderstanding, high word overlap, and attribution errors. Addressing the top two error categories, external knowledge and numerical reasoning errors, will require broader pre-training on encyclopedic and reasoning datasets before fine-tuning for NLI-specific tasks.

2.5 Fine-Tuning

To enhance the model's performance, we fine-tuned it in multiple stages, focusing on specific error types identified during evaluation. Starting with the XNLI-trained model, we fine-tuned on the metaphor training set described earlier for five epochs. Performance on the validation set peaked at epoch 4, and this checkpoint was used for subsequent fine-tuning.

For the three additional error types — negation misunderstanding, attribution errors, and high word overlap — identified from the ANLI evaluation, we used a pre-translated version of ANLI available on HuggingFace[1]. This dataset comprises of multiple machine translated NLI datasets, but we focused exclusively on ANLI. Custom scripts were developed to extract examples corresponding to each error category:

- **Negation**: Records where the hypothesis contained negation words (e.g., ”not”, ”never”).
- **Attribution**: Records containing attribution phrases (e.g., ”owned by”, ”performed by”).
- **High Word Overlap**: Records where the cosine similarity between the premise and hypothesis was at least 0.8.

This extraction process was run on the English dataset and then pulled in both English and Russian records yielding 34,540 negation examples, 28,836 attribution examples, and 3,236 high word overlap examples. To avoid over-representing ANLI-derived examples compared to the smaller metaphor dataset (2,403 examples), we fine-tuned the model for one epoch on each of these datasets individually.

While this sequential fine-tuning improved performance on ANLI (see Table 1), we observed a degradation in performance on the metaphor dataset, indicating catastrophic forgetting (Kirkpatrick et al. 2017) as the model adapted to ANLI's structure. To address this, we created a combined fine-tuning dataset that merged all four datasets. The metaphor data was oversampled (included four times) to reflect the four epochs initially used for its fine-tuning and to balance the smaller metaphor dataset against the larger ANLI-derived datasets. The XNLI-trained model

[1] https://habr.com/ru/articles/582620/

Training	**English**				**Russian**				**Total**			
	X	M	A	M+A	X	M	A	M+A	X	M	A	M+A
XNLI	79.5	61.3	29.3	31.6	72.1	49.7	30.3	31.6	75.8	55.5	29.8	31.6
+Metaphor	78.1	73.4	28.4	31.6	69.7	60.3	29.7	31.9	74.0	66.8	29.1	31.8
+Negation	74.4	60.7	35.3	37.3	66.1	39.9	35.1	35.5	70.2	50.1	35.2	36.4
+Attribution	69.9	52.7	41.1	42.0	61.7	40.9	38.9	39.1	65.9	46.8	40.0	40.6
+High overlap	71.5	55.4	40.5	41.6	62.2	45.4	39.8	40.1	66.9	50.4	40.2	40.9
Full fine-tune	76.4	69.8	40.1	42.3	68.8	57.2	39.5	41.1	72.6	63.3	39.8	41.7

Table 1: Model Performance Across Test Sets for English, Russian, and Combined Languages. Columns represent test sets: ‘X’ refers to XNLI, ‘M’ to Metaphor dataset, and ‘A’ to ANLI Curated Error dataset. Metric measured is the F1 score.

was then fine-tuned on this combined dataset for one epoch.

3 Evaluation

To assess the impact of fine-tuning on both XNLI and curated datasets, we evaluated the model across four test sets: XNLI (X), the metaphor dataset (M), the ANLI curated error dataset (A), and a combined challenging test set (M+A). Table 1 presents the F1-scores for each test set across English, Russian, and the combined total.

Fine-tuning on the metaphor dataset (+Metaphor) yielded significant improvements in F1-scores for metaphors, with an increase of 12.1 points in English (from 61.3 to 73.4) and 10.6 points in Russian (from 49.7 to 60.3). These results highlight the effectiveness of targeted fine-tuning in addressing specific linguistic challenges, such as metaphors. However, these gains did not generalize to the ANLI dataset, where performance remained largely unchanged, emphasizing the domain-specific nature of metaphor fine-tuning.

Subsequent fine-tuning on curated datasets for negation misunderstanding (+Negation), attribution errors (+Attribution), and high word overlap (+High Overlap) resulted in notable performance gains on the ANLI dataset:

- Negation: +6.1 (29.1 → 35.2)
- Attribution: +4.8 (35.2 → 40.0)
- High Overlap: +0.2 (40.0 → 40.2)

These results demonstrate the value of error-specific fine-tuning in improving robustness to adversarial challenges. Notably, the increase in the Russian dataset with the high word overlap training data indicates that shuffled word order is an especially relevant area for languages with a flexible grammar structures like Russian. Unlike English, where word order typically follows the canonical subject-verb-predicate pattern, Russian relies more on word endings to convey grammatical relationships. This allows sentences to retain their meaning even when word order is varied.

For example, the Russian sentences:

- "Мальчик читает книгу." (”The boy reads a book.”)
- "Мальчик книгу читает." (”The boy a book reads.”)
- "Книгу читает мальчик." (”A book is read by the boy.”)
- "Книгу мальчик читает." (”A book by the boy is read.”)
- "Читает мальчик книгу." (”Reading by the boy a book.”)
- "Читает книгу мальчик." (”Reading a book by the boy.”)

have identical meanings and are all valid sentences despite differing word orders (unlike the English translations which are not all valid). Including training examples that maintain high word overlap but vary in word order could further improve the model’s ability to generalize these grammatical rules, enhancing its performance on Russian datasets.

The slight decline in metaphor performance after fine-tuning for ANLI suggests over-fitting to linguistic structures found in ANLI, reinforcing the need for a balanced fine-tuning strategy. The final stage of training involved combining all fine-tuning datasets (metaphors, negation, attribution, and high overlap) into a single dataset. This combined fine-tuning approach resulted in more balanced improvements across all test sets, achieving overall F1 scores of 72.6 on XNLI, 63.3 on metaphors, 39.8 on ANLI, and 41.7 on the combined test set (M+A). While gains on metaphors and ANLI were a bit more modest compared to individual fine-tuning, this approach mitigated the effects of catastrophic forgetting observed in sequential training. Catastrophic forgetting refers to the phenomenon where a model loses performance on previously learned tasks when fine-tuned on new tasks or datasets. In this research, we observed this when the model's performance on the metaphor dataset declined after fine-tuning on ANLI-derived datasets. This effect leads to trade-offs between improving one dataset and degrading another.

4 Discussion and Future Work

While fine-tuning on specific error types improved performance, significant challenges remain in addressing errors requiring external knowledge and numerical reasoning. Our analysis highlights a few promising directions for future research:

4.1 Numerical Reasoning

Our analysis suggests incorporating broader pre-training on numerical reasoning datasets before NLI-specific training. The AQuA-RAT dataset Ling et al. 2017, which focuses on algebraic word problems with rationales, represents one potential resource. However, to avoid catastrophic forgetting of NLI tasks, we recommend a staged approach:

- Initial pre-training on numerical reasoning datasets
- Intermediate training on multilingual NLI tasks
- Final fine-tuning on NLI-specific examples to address error areas

Future work should focus on creating specialized NLI datasets that specifically target numerical reasoning capabilities:

- Collecting premise-hypothesis pairs involving quantitative comparisons
- Creating examples that require basic arithmetic operations
- Creating examples that substitute numbers for words representing the numerals

4.2 External Knowledge Integration

For our knowledge enhancement strategy we recommend pre-finetuning on Wikipedia and other encyclopedic sources, with special attention to factual relationships and commonsense reasoning. Once a model has expanded its knowledge we would then finetune on examples that require multi-reasoning steps where first the model must make a connection to its known internal knowledge and then connect a hypothesis to a stated premise. One example of this is the following where a model must first know that in South Africa cars are driven on the left hand side and then apply this knowledge to tie together the hypothesis to the premise:

- Premise: The Volkswagen Citi Golf was a car produced by Volkswagen in South Africa from 1984 until 21 August 2009....
- Hypothesis: The Volkswagen Citi Golf was designed to drive on the left side of the road.

4.3 Conclusion

This study highlights the complexity and challenges of multilingual NLI, particularly when addressing diverse linguistic phenomena such as metaphors, negation, and high word overlap. By fine-tuning on targeted error datasets, we demonstrated significant improvements in robustness and came across challenges like catastrophic forgetting and lower generalization to adversarial datasets. Our findings underscore the need for incorporating diverse datasets and advanced strategies like sequential training and multi-task learning.

Future advancements in multilingual NLI should address knowledge integration and numerical reasoning through broader pre-training and fine-tuning pipelines. Ultimately, these efforts can lead to models capable of more nuanced understanding and robust performance across languages, bringing us closer to achieving truly generalizable natural language inference frameworks.

References

Conneau, Alexis et al. (2018). "XNLI: Evaluating Cross-lingual Sentence Representations." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, pp. 2475–2485. URL: https://arxiv.org/pdf/1809.05053.pdf.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186. URL: https://arxiv.org/pdf/1810.04805.pdf.

Kirkpatrick, James et al. (2017). "Overcoming Catastrophic Forgetting in Neural Networks." In: *Proceedings of the National Academy of Sciences* 114.13, pp. 3521–3526. DOI: 10.1073/pnas.1611835114.

Ling, Wang et al. (2017). "Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 158–167. DOI: 10.18653/v1/P17-1015. URL: https://aclanthology.org/P17-1015.

Mohler, Michael et al. (2016). "Introducing the LCC Metaphor Datasets." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portoro, Slovenia, pp. 4221–4227. URL: https://aclanthology.org/L16-1668.pdf.

Nie, Yixin et al. (2020). "Adversarial NLI: A New Benchmark for Natural Language Understanding." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901. DOI: 10.18653/v1/2020.acl-main.441. URL: https://www.aclweb.org/anthology/2020.acl-main.441.