The Impact of Synthetic Data on Text Summarization Quality

Problem

Russian text summarization models face challenges due to a scarcity of high-quality, diverse labeled datasets, limiting their performance.

Method

We will analyze the optimal ratio of synthetic to real data for diverse domains, assess the impact of semantic filtering, and evaluate model adaptation across architectures and domains.

Contribution

Design and implementation of synthetic data generation techniques and analysis of their impact on summarization quality.

Data Description

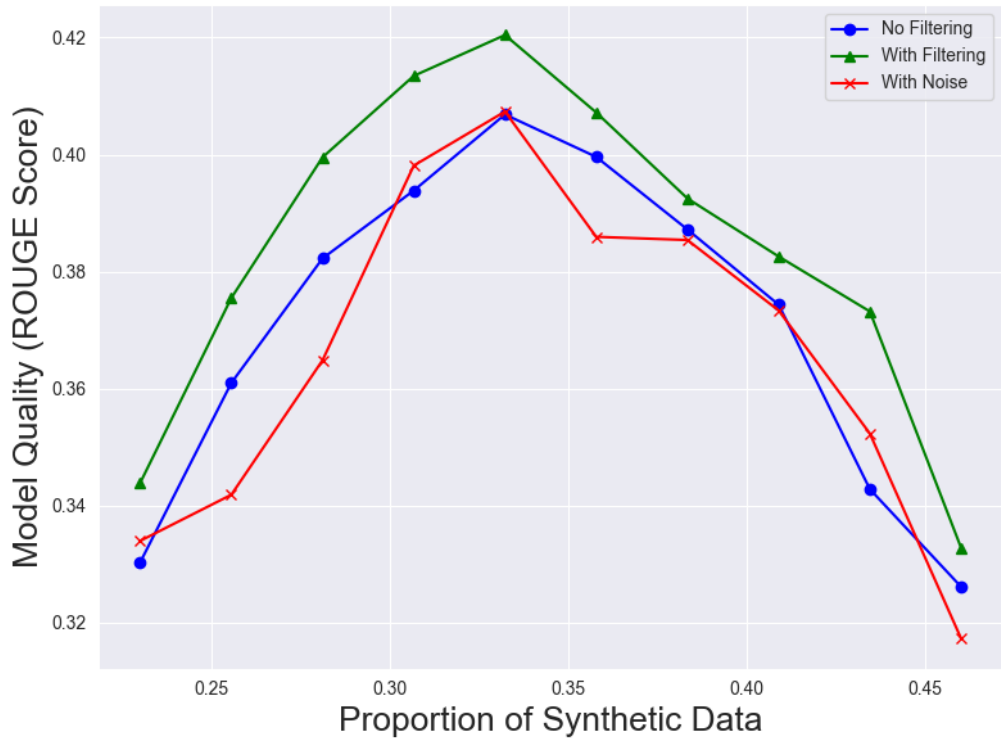
Datasets:

- Real data:
  Gazeta, WikiLingua, Matreshka, DialogSum, etc.
  ≈200k samples in Russian without long context, covering several domains.
- Synthetic data:
  Automatically generated summaries to complement existing datasets.
  Expected Benefits: Enhanced performance by eliminating domain gaps and enriching linguistic diversity, evaluated using both classical (e.g., ROUGE, BLEU) and advanced metrics (e.g., BERT Score).

Error Analysis and Expected Plots

Analysis Plan:

- **Impact of Synthetic Data and Semantic Filtering:** Compare model metrics with varying proportions of synthetic data and evaluate the effect of semantic filtering.
- **Noise Impact:** Assess models stability by adding noise to synthetic data.
- **Learning Curve:** Track metrics changes over iterations with various data compositions.



The impact of synthetic data, noise, and filtering on quality.