SOVEREIGN: What is the impact of back-translation paraphrasing techniques on QA model generalization across different mod

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

To produce a domain-agnostic question answering model for the Machine Reading Question Answering (MRQA) 2019 Shared Task, we investigate the relative benefits of large pre-trained language models, various data sampling strategies, as well as query and context paraphrases generated by back-translation. We find a simple negative sampling technique to be particularly effective, even though it is typically used for datasets that include unanswerable questions, such as SQuAD 2.0. When applied in conjunction with per-domain sampling, our XLNet (Yang et al., 2019)-based submission achieved the second

1 Introduction

Analysis of: An Exploration of Data Augmentation and Sampling Techniques for Domain-Agnostic Question Answering. Research goal: What is the impact of back-translation paraphrasing techniques on QA model generalization across different model scales and sampling strategies in domain-agnostic question answering tasks?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

3 papers retrieved. 13 claims extracted, 0 verified. Tribunal: 5.3/10 → REJECT (revision_round=1). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
The checkpoint with the highest Out-Domain Macro-Average was selected as the best for that training run.	×	0.03
Our multi-domain dataset originally consisted of 75k examples from every training set.	×	0.01
We modified this to a maximum of 120k samples from each dataset, 100k from SearchQA, and using only one detected answer	×	0.00
The improvement is exaggerated at the shorter max sequence length (MSL) of 200, where including NA segments increases Ou	×	0.02
SearchQA is the largest dataset by number of examples.	×	0.02
SearchQA generated 657K segments, double that of the next largest dataset.	×	0.02
We found this drastically outperformed the typical practice of excluding these segments.	×	0.00
Models with multi-domain pre-fine-tuning NewsQA, SearchQA, and TriviaQA performed the worst on the out-domain Macro-Aver	×	0.04
SQuAD fine-tuned model achieves the best results on both in and out-domain 'Macro-Average' Exact Match.	×	0.08
The SQuAD dataset contains fine-tuned models that perform best on both in and out-domain 'Macro-Average' Exact Match.	×	0.08
The more sophisticated techniques including back-translated augmentations yield no noticeable improvement.	×	0.02
Much simpler techniques offer significant improvements.	×	0.09
Negative samples designed to teach the model when to abstain from predictions prove highly effective out-domain.	×	0.05

References

- http://arxiv.org/abs/2509.07471v2
- http://arxiv.org/abs/2312.00912v1

- http://arxiv.org/abs/1912.02145v1