Cross-Lingual Transferability of Zero-Shot Question Generation Versus Fine-Tuned Retrieval on XQuAD and MLQA

Assignee Research

June 12, 2026

Abstract

Question answering (QA) models have shown rapid progress enabled by the availability of large, high-quality benchmark datasets. Such annotated datasets are difficult and costly to collect, and rarely exist in languages other than English, making training QA systems in other languages challenging. An alternative to building large monolingual training datasets is to develop cross-lingual systems which can transfer to a target language without requiring training data in that language. In order to develop such systems, it is crucial to invest in high quality multilingual evaluation benchmarks to m

1 Introduction

This paper examines: MLQA: Evaluating Cross-lingual Extractive Question Answering. Research question: How does the cross-lingual transferability of zero-shot question generation models compare to fine-tuned retrieval models on XQuAD and MLQA benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

16 papers retrieved. 19 claims extracted; 12 independently verified. Quality review score: 7.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Manually translating documents at sufficient scale entails huge translator workloads and could result in unnatural docum	✓	0.18
Exploiting existing naturally-parallel documents provides high-quality documents without requiring manual translation.	✓	0.22
Cross-lingual understanding benchmarks are typically based on classification.	✓	0.15
Extracting spans in different languages represents a different language understanding challenge than classification.	✓	0.22
Most existing extractive QA datasets in various languages were created at different times by different authors with diff	✓	0.23
Wikipedia enables the collection of data in many diverse languages at scale due to its size and multi-linguality.	✓	0.19
Wikipedia has been used to build many existing QA training resources.	✓	0.21
English has the largest Wikipedia.	×	0.09
The LASER toolkit achieves state-of-the-art performance in mining parallel sentences.	✓	0.21
LASER uses multilingual sentence embeddings and a distance or margin criterion in the embeddings space to detect paralle	✓	0.22
Starting with 5.4M parallel English/German sentences, the number of N-way parallel sentences quickly decreases as more l	✓	0.26
7-way parallel sentences lack linguistic diversity and often appear in the first sentence or paragraph of articles.	✓	0.24
The dataset contains 5.4M parallel sentences for the de-en pair.	×	0.09
The dataset contains 1.1M parallel sentences for the es-en pair.	×	0.07
The dataset contains 83.7k parallel sentences for the ar-en pair.	×	0.10
The dataset contains 24.1k parallel sentences for the zh-en pair.	×	0.10
The dataset contains 9.2k parallel sentences for the vi-en pair.	×	0.07
The dataset contains 1,340 parallel sentences for the hi-en pair.	×	0.06
The study uses sentences that are 4-way parallel as a compromise between language-parallelism and the number and diversi	✓	0.20

References

- http://arxiv.org/abs/1910.07475v3
- http://arxiv.org/abs/2408.11942v1
- http://arxiv.org/abs/2205.02303v1