Sequence-to-Sequence Pre-Training Effects on Bidirectional Encoding in GLUE Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: To what extent does the sequence-to-sequence pre-training objective in unified models degrade bidirectional encoding accuracy on semantic textual similarity tasks within GLUE. 5 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Unified Language Model Pre-training for Natural Language Understanding and Generation. Research question: To what extent does the sequence-to-sequence pre-training objective in unified models degrade bidirectional encoding accuracy on semantic textual similarity tasks within GLUE?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.3/10.

3 Results

10 papers retrieved. 5 claims extracted; 0 independently verified. Quality review score: 6.3/10.

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
UNILM achieves a ROUGE-1 score of 43.33, ROUGE-2 score of 20.21, and ROUGE-L score of 40.51 on the CNN/DailyMail summari	×	0.12
UNILM achieves a ROUGE-1 score of 38.45, ROUGE-2 score of 19.45, and ROUGE-L score of 35.75 on the Gigaword abstractive	×	0.13
UNILM achieves a ROUGE-1 score of 32.96, ROUGE-2 score of 14.68, and ROUGE-L score of 30.56 on the Gigaword abstractive	×	0.10
UNILM outperforms the best system in the DSTC7 shared task across all evaluation metrics.	×	0.04
UNILM achieves a score of 61.1 on CoLA, 94.5 on SST-2, 90.0 on MRPC, 87.7 on STS-B, 71.7 on QQP, 87.0/85.9 on MNLI-m/mm,	×	0.05

References

- `http://arxiv.org/abs/1905.03197v3`
- `http://arxiv.org/abs/1909.11059v3`
- `http://arxiv.org/abs/2209.15329v3`