PoSum-Bench : un benchmark pour l'évaluation du biais positionnel dans la synthèse conversationnelle

Xu Sun[1, 2] Lionel Delphin-Poulat[1] Christèle Tarnec[1] Anastasia Shimorina[1]
(1) Orange Research
(2) Université Paris Cité, Paris, France
{xu.sun, lionel.delphin-poulat, christele.tarnec,
anastasia.shimorina}@orange.com

RÉSUMÉ

Les grands modèles de langue (LLMs) sont de plus en plus utilisés pour la synthèse de conversations en zero-shot, mais présentent souvent un biais positionnel, tendant à surreprésenter le contenu situé au début ou à la fin d'une conversation au détriment du milieu. Pour répondre à ce problème, nous introduisons PoSum-Bench, un benchmark complet pour l'évaluation du biais positionnel dans la synthèse conversationnelle, comprenant des jeux de données conversationnelles diversifiés en anglais et en français couvrant des réunions formelles, des conversations informelles et des interactions de service client. Nous proposons une métrique originale au niveau des phrases, fondée sur la similarité sémantique, permettant de quantifier la direction et l'amplitude du biais positionnel dans les résumés générés, offrant ainsi une évaluation systématique et sans référence selon les positions dans la conversation, les langues et les contextes conversationnels.

ABSTRACT

PoSum-Bench : A Benchmark for Evaluating Positional Bias in Conversational Summarization

Large language models (LLMs) are increasingly used for zero-shot conversation summarization, but often exhibit positional bias—tending to overemphasize content from the beginning or end of a conversation while neglecting the middle. To address this issue, we introduce PoSum-Bench, a comprehensive benchmark for evaluating positional bias in conversational summarization, featuring diverse English and French conversational datasets spanning formal meetings, casual conversations, and customer service interactions. We propose a novel semantic similarity-based sentence-level metric to quantify the direction and magnitude of positional bias in model-generated summaries, enabling systematic and reference-free evaluation across conversation positions, languages, and conversational contexts.

MOTS-CLÉS : biais positionnel, synthèse conversationnelle, évaluation, grands modèles de langue, benchmark.

KEYWORDS: positional bias, conversational summarization, evaluation, large language models, benchmark.

ARTICLE ACCEPTÉ À : Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP).
URL : https://aclanthology.org/2025.emnlp-main.404/