ContraDoc: Understanding Self-Contradictions in Documents with Large Language Models

Jierui Li¹, Vipul Raheja², Dhruv Kumar²

1. The University of Texas at Austin 2. Grammarly

Introduction

Motivation: A text is considered self-contradictory when it contains multiple ideas or statements that inherently conflict. Humans struggle to identify contradictions in unfamiliar, informative texts, particularly when contradictions are widely separated in long documents, underscoring the need for automated text analysis tools.

ContraDoc: Understanding Self-Contradictions

Machine-Human Collaboration:

a. Use LLM to find statements in the document and generate contradictory statements.

b. Inserting the contradictory statement or replacing the original statement with it based on automatic metrics.

c. Human annotators and expert filter & tag the candidate self-contradictory documents.

Evaluation Metrics

Detect

If there's self-contradiction

Top-k Contradictions: Point out evidence for self-contradiction

Top-k Contradictions: Point out evidence for self-contradiction

Judge then Find: First Judge, then Find (Point-Out)

• Determine one doc is self-contradictory or not. If is, give 2 self-contradictory sentences or 1 if (the self-contradiction is within a sentence). It’s considered correct only when the model did answer “yes” and provide correct evidence.

• Evaluate on 449 positive articles and 445 negative articles from similar distribution.

Experiments

Binary Judgment experiments show that while GPTs and PaLM are under-performing (tend to predict “no”), LLaMA is over-performing (tend to predict “yes”) to SC.

Further experiments in self-contradiction top 5/judge-then-find show that even when LLaMA is answering “yes”, it doesn’t seem to find where the self-contradiction lies.

Performance metrics are available at: https://github.com/ddhruvkr/CONTRADOC

Evaluate Metrics & Dataset are available at: https://github.com/ddhruvkr/CONTRADOC