CS 6120: Fact Checker

Ashwin Satish, Rohan Chouhan, Daniel Xiong

Abstract

We propose to build a fact-checking system that verifies claims by retrieving and reasoning over multiple pieces of evidence. Given a claim, our system will retrieve relevant documents, identify key evidence passages, and combine them through multi-hop reasoning to classify claims as SUPPORTED, REFUTED, or NOT_ENOUGH_INFO.

1 Introduction

Fact-checking has become increasingly critical in the age of misinformation. Traditional fact-checking systems often rely on single pieces of evidence, but many real-world claims require reasoning across multiple sources. This project addresses multi-hop fact verification, where claims must be validated by connecting 2-4 pieces of evidence. We aim to build an end-to-end pipeline combining retrieval and verification components to classify claims as SUPPORTED, REFUTED, or NOT_ENOUGH_INFO. Our motivation is to advance automated fact-checking capabilities for complex claims that require chaining evidence from multiple documents.

2 Background/Related Work

Our work builds on several key advances in fact-checking and multi-hop reasoning. Thorne et al. (2018) introduced FEVER, establishing baseline pipelines for evidence retrieval and verification. Jiang et al. (2020) created HoVer, our primary dataset designed specifically for multi-hop reasoning with 2-4 evidence chains. Zhou et al. (2019) proposed GEAR, using graph attention networks to model evidence relationships. Zhao et al. (2020) demonstrated transformer architectures for multi-hop reasoning through compositional information aggregation. Xiong et al. (2021) advanced multi-hop dense retrieval techniques. Guo et al. (2022) provided a comprehensive survey of automated fact-checking methods and challenges.

3 Data

Describe the datasets and any preprocessing, cleaning that you may have done.
Hover dataset:

1. 18171 Train Entries, 9.2 MB
   a. 11023 SUPPORTED
   b. 7148 NOT_SUPPORTED
2. 4000 Dev Entries, 2.2 MB
   a. 2000 SUPPORTED
   b. 2000 NOT_SUPPORTED
3. 4000 Test Entries, 899 KB
   a. No labels given
4. This data is all claims, as well as the True/False labels for whether they are factually accurate or not

HotPotQA Wikipedia Dump:

1. 15519 already preprocessed Wikipedia articles in the form of JSONs

4 Methods

Our pipeline has two stages: (1) Retrieval - BM25 baseline via Elasticsearch retrieving top-20 passages, Sentence-BERT dense retrieval (all-MiniLM-L6-v2) with FAISS indexing, and cross-encoder BERT reranker for top-5 selection. (2) Verification - Fine-tuned RoBERTa-base for 3-way classification with input format: [CLS] claim [SEP] evidence_1 [SEP] evidence_2 [SEP] evidence_3.

5 Experiments

Describe your experiments. How were the models trained, which hyperparameters, training and inference approaches, etc..

5.1 Results

We've downloaded the Wikipedia dataset and are currently working on indexing it so that it can be used for fact checking