Fact and scientific claim verification with Natural Language Understanding and reasoning

Alan Song
amsong@andrew.cmu.edu

Meghana Balarama
mbalaram@andrew.cmu.edu

Abstract

The spread of misinformation and disinformation over social media is a problem of growing concern, necessitating the use of automatic fact checking and fact verification about claims on vaccines to climate change. Fact-checked false claims can still spread on social media and thus it is also important to identify previously fact-checked claims. BERT classifier models fail to respond to changes in compositional and lexical meaning and hence the ability of such fact checking methods need to be tested on their ability to reason about real-world facts. In this project, we propose to study and improve the reasoning ability of fact verification models over structured and unstructured information by finding precise evidence to support or dismiss a claim.

1 Introduction

The goal of fact verification is to validate a claim in the context of evidence - given a claim C and evidence P as inputs, it predicts if the claim is supported, refuted, or can not be verified by the information in P.

The popular publicly available datasets for this task are FEVER (Fact Extraction and VERification) ([7]), FEVEROUS (Fact Extraction and VERification Over Unstructured and Structured information) ([1]), VitaminC (VitaminC Robust Fact Verification) ([6]), and CLIMATE-FEVER ([2]) each containing verified claims. Each claim is annotated with evidence in the form of sentences and/or cells from tables in Wikipedia, as well as a label indicating whether this evidence supports, refutes, or does not provide enough information to reach a verdict. However, large scale training data is not available for every new domain and human annotation of claims is expensive, thus necessitating the automatic generation of training data.

The existing models for fact verification propose pipeline models of abstract retrieval, rationale selection and stance prediction. Such works have the problems of error propagation among the modules in the pipeline and lack of sharing valuable information among modules [11]. Certain other methods use an ensemble of pre-trained models for multi-model fact verification to achieve better performance [10]. To successfully distinguish false articles from genuine ones, a model must be able to incorporate world knowledge into its computation, along with being proficient in natural language understanding. Models that detect false information can be fooled by carefully tweaked input and hence such a detector should base its classification on the semantic content of its input and its relation to real-world facts, and not on superficial features of the text [3]. The models developed for fact verification should be tested for their ability to reason about real-world facts - they should test "understanding": compositional semantics, lexical relations, and sensitivity to modifiers.

Another crucial aspect of fact verification is computing precise evidences - minimal sets of sentences that support or refute a given claim, rather than larger evidences - since larger evidences may contain conflicting pieces some of which support the claim while the other refute, thereby misleading the model. [9] proposes a DQN-based approach to retrieval of precise evidences and demonstrates improvements in achieving accurate claim verification.

2 Literature Review

Creating a dataset by manually writing claims and linking them to their supportive evidence for every new domain is expensive. To alleviate the need to create human-annotated training data, [5] proposes a method to automatically generate large-scale (evidence, claim) pairs to train the fact verification model. The proposed model Question Answering for Claim Generation (QACG) generates three types of claims from any given evidence - claims that are supported by the evidence, claims that are refuted by the evidence, claims that can not be verified because there isn't enough evidence available. The claim generation is based on Question Generation which automatically asks questions from textual inputs. A Question Generator generates a question–answer pair (Q;A) for the evidence which is then converted into a claim C (QA-to-Claim) based on the following logical assumptions: a) if P can answer Q and A is the correct answer, then C is a supported claim; b) if P can answer Q but A is an incorrect answer, then C is a refuted claim; c) if P cannot answer Q, then C is a NEI claim. Crowd sourced human annotated datasets also suffer from undesired biases in the data that cause models to learn spurious patterns. [4] proposes a two-stage contrastive data-augmentation pipeline to generate new claims and evidences from existing samples that debiases fact verification models. Cross-wise pairing of the generated samples with the original pair to form contrastive samples helps the model to rely less on spurious patterns and learn more robust representations.

A multi-modal fact verification network is proposed in [10], which extracts features from text and images using two pre-trained models and fuses different modalities and same modality but different sources using multiple co-attention networks. This model model achieves a competitive performance without using auxiliary tasks or extra information. [11] aims to solve the error propagation problem and the lack of information sharing among modules with their ARSJoint model which jointly learns the three models for abstract retrieval, rationale selection and stance prediction. It is based on Machine Reading Comprehension (MRC) framework and learns additional information using the claim content as query. It uses BioBERT and RoBERTa as its backbone networks. [8] performs fact-checking using a hierarchical multi-head attentive network that jointly combines multi-head word-level attention and multihead document-level attention to aid explanation in both word-level and evidence level.

Cleverly written fake articles can fool fact-verification systems and evade detection. A model should be able to incorporate world knowledge into its computation in addition to natiral language understanding to successfully distinguish fake articles from genuine ones. The model should base its classification on the semantic content of its input and its relation to real-world facts, and not on superficial features of the text. [3] proposes an adversarial benchmark that targets three aspects of a model's "understanding" - whether it has the ability to employ semantic composition, whether it incorporates world knowledge of political parties, and whether adverb intensity is employed as a signal of fake news.

3 Gaps identified

Large scale training data is not available for every new domain and human annotation of claims is expensive, thus necessitating the automatic generation of training data. Models that detect false information can be fooled by carefully tweaked input and hence such a detector should base its classification on the semantic content of its input and its relation to real-world facts, and not on superficial features of the text. These models need to be tested for their ability to understand and reason about real-world facts and their robustness to adversarial attacks.

4 Proposed Approach

We propose:

(a) Automatically generating training data and data augmentations for domains for which large-scale training data is not available. We are yet to choose a domain, but at this time we are considering climate-change claims as a potential domain.

(b) Multi-task training of a joint model or fine-tuning an ensemble of pre-trained models on this mix of real-world synthetic data for fact verification, along with computing precise evidences

(c) Adversarial testing of the trained model to identify the vulnerabilities of the model.

5 Expected Outcomes

We expect that our proposed approach will improve the accuracy of claim verification and help in analysing the vulnerabilities of the model to adversarial attacks that are meant to fool the model. If we are able to achieve good performance on claim verification by augmenting a small dataset with automatically generated data and data augmentations, this would also lead to a framework that could then possibly be extended to multiple other domains.

References

[1] Rami Aly et al. "FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information". In: *CoRR* abs/2106.05707 (2021). arXiv: 2106.05707. URL: https://arxiv.org/abs/2106.05707.

[2] Thomas Diggelmann et al. "CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims". In: *CoRR* abs/2012.00614 (2020). arXiv: 2012.00614. URL: https://arxiv.org/abs/2012.00614.

[3] Lorenzo Jaime Yu Flores and Yiding Hao. *An Adversarial Benchmark for Fake News Detection Models*. 2022. arXiv: 2201.00912 [cs.CL].

[4] Minwoo Lee et al. "CrossAug: A Contrastive Data Augmentation Method for Debiasing Fact Verification Models". In: *CoRR* abs/2109.15107 (2021). arXiv: 2109.15107. URL: https://arxiv.org/abs/2109.15107.

[5] Liangming Pan et al. "Zero-shot Fact Verification by Claim Generation". In: *CoRR* abs/2105.14682 (2021). arXiv: 2105.14682. URL: https://arxiv.org/abs/2105.14682.

[6] Tal Schuster, Adam Fisch, and Regina Barzilay. "Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence". In: *CoRR* abs/2103.08541 (2021). arXiv: 2103.08541. URL: https://arxiv.org/abs/2103.08541.

[7] James Thorne et al. "FEVER: a large-scale dataset for Fact Extraction and VERification". In: *CoRR* abs/1803.05355 (2018). arXiv: 1803.05355. URL: http://arxiv.org/abs/1803.05355.

[8] Nguyen Vo and Kyumin Lee. *Hierarchical Multi-head Attentive Network for Evidence-aware Fake News Detection*. 2021. arXiv: 2102.02680 [cs.AI].

[9] Hai Wan et al. "A DQN-based Approach to Finding Precise Evidences for Fact Verification". In: Jan. 2021, pp. 1030–1039. DOI: 10.18653/v1/2021.acl-long.83.

[10] Wei-Yao Wang and Wen-Chih Peng. *Team Yao at Factify 2022: Utilizing Pre-trained Models and Co-attention Networks for Multi-Modal Fact Verification*. 2022. arXiv: 2201.11664 [cs.CV].

[11] Zhiwei Zhang et al. "Abstract, Rationale, Stance: A Joint Model for Scientific Claim Verification". In: *CoRR* abs/2110.15116 (2021). arXiv: 2110.15116. URL: https://arxiv.org/abs/2110.15116.