**Automated Fact Checking: Task formulations, methods and future directions[[1]](#footnote-0)**

Summary

* Fact checking ranges from 1 hour to a few days
* Journalism defined by scholars as “discipline of verification” to separate it from “entertainment, propaganda, fiction, or art”.
* Inputs
  + Subject-predicate-object triples (eg. London, capital\_of, UK)
    - Require non-trivial processing to convert claims to triples
  + Textual claims (short sentences)
    - Popular due to availability of fact checked claims
    - Taxonomy of claims (HeroX fact checking challenge)
      * Numerical claims - numerical properties about entities
      * Entity and event properties - eg. professional qualifications
      * Position statements - eg. political entity support to policy
      * Quote verification - whether claim states source of quote & context
  + Entire document as input
    - Identify claims then fact check
    - High complexity required to extract claims in triples by relation extraction or supervised sentence-level classification
* Sources of evidence
  + No factual evidence
    - Surface-level linguistic features in claims
    - Metadata - originator of claim, speaker profile, media source
    - Context acts as prior to improve classification accuracy
  + Knowledge graphs
    - Identify/retrieve element supporting or refuting claim
      * Eg. Identify SPO triple, compute truth label by error between claim values and retrieved values
      * Key limitation - Assumes true facts relevant to claim are present
    - Graph topology
      * If fact unlikely to occur, does not negate its truthfulness
      * Improbable but believable claims more likely to go viral
  + Text (eg. encyclopedia, policy docs, verified news, scientific journals)
    - Article headlines
    - Entire documents
    - Wikipedia
    - Repository of fact checked claims
      * Prediction of veracity labels
      * Limited to existing fact checked claims
  + Social network distribution - crowd behaviour
    - Useful where textual & structured knowledge bases unavailable
  + Trustworthiness of sources (separated task)
* Output
  + Binary classification task (true/ false)
  + Degree of truthfulness (true, mostly-true, half-true, etc)
    - Reasoning is complex, sometimes inconsistent, difficult to express
  + Supported, refuted or reported by article headline, or irrelevant (incomplete)
  + Supported/refuted by Wikipedia, or insufficient information
* Fact Checking Datasets
  + Vlachos and Riedel (2014) 221 Politifact and Channel4 claims, speaker, hyperlink
  + Wang (2017) 12.8K Politifact claims, meta-data (speaker affiliation), context (speech, tweet, op-ed piece)
    - No machine-readable evidence beyond originator metadata
  + Rashkin et al. (2017) 74K Politifact claims (hoax, satire, propaganda, trusted)
    - Fact check using linguistic features but not evidence
  + Ferreira and Vlachos (2016) 300 Claims, 2595 news articles stances (for, against, observing claim)
  + HeroX Fast & Furious Fact Checking Challenge (2016) 90 (41 practice, 49 test) labeled claims
    - Insufficient size of dataset, evidence not provided
  + Thorne et al. (2018) 185K claims, properties of entities, concepts, to be verified by Wikipedia articles
    - Allow for evidence selection subtask
* Methods
  + Supervised
    - Requires additional world knowledge, not provided with claim
    - Detect fact-check worthy claims
  + Network analysis
    - Predict if unobserved triple like to appear in graph as path ranking problem
      * Truth verdict derived from cost of traversing path
      * Compare linguistic features of triple with other triples in same topic
  + Recognizing Textual Entailment (RTE)
    - Assume textual evidence given
  + Distantly supervised relation extraction
    - Identify surface patterns between 2 entities in a knowledge graph
    - Comparison of values but many different relations to be distinguished
  + Matching claim with existing, previously fact checked claims
    - Sentence-level textual similarity
    - Only for repeated/ paraphrased claims
  + Speaker profiling (credit history of originator)
    - May not have recorded history
    - Strong reliance has ethical implications
* Related Tasks
  + Verification
    - Must consider information of unknown veracity as evidence
    - Predict authoritativeness of web pages
      * PageRank - hyperlink topology
      * TrustRank - graph-connectedness to known-bad nodes
    - Knowledge-based Trustworthiness scoring
      * Predict fact accuracy by fact extraction method & website
  + Common Sense Reasoning
    - Warrant (rationale of argument) required to support claim from given premise. Habernal et al. (2018) task & dataset
    - Extract common sense knowledge from WordNet using Natural Logic
  + Subjectivity and Emotive Language
    - Supervised classifiers augmented with lexicons (LIWC, sentiment, hedging, dramatic language)
  + Deceptive Language Detection
    - Syntactic style rather than word-based content
    - Depends on context or else cannot be used to fact check alone
  + Rumor Detection
    - By language subjectivity and growth of readership
    - Less correlated with true/false
  + Speaker Profiling
    - Credit history does not account for specific topics originator lies about
    - Compatibility of claim with originator’s profile
    - Author characteristics (age, gender) influence linguistic choices
  + Click Bait Detection
    - Positive results from simple linguistic features in headline without evidence
* Open Research Challenges
  + Open-world knowledge
    - Address information retrieval challenge and its evaluation
    - Consider verification jointly with fact checking
    - Text-based claim but evidence in other modalities (images, videos)
* Complexity of fact checking by journalists is currently beyond abilities of systems due to complex reasoning needed

1. "Automated Fact Checking: Task formulations, methods and future ...." 20 Jun. 2018, <https://arxiv.org/abs/1806.07687>. Accessed 7 Dec. 2018. [↑](#footnote-ref-0)