**FEVER: a large-scale dataset for Fact Extraction and VERification**

**Summary**

* Verification against textual sources
* Generate 185,445 claims by altering sentences extracted from Wikipedia
* Verify claims without knowledge of the original extracted sentence
* Claims classified as Supported, Refuted or NotEnoughInfo by annotators achieved 0.6841 Fleiss k in claim verification classification
  + Fleiss’ kappa
    - Statistical measure to assess reliability of agreement between fixed no. of raters when assigning categorical ratings to a no. of items or classifying items.
    - Calculates degree of agreement in classification over that expected by chance.
    - “Measure of how consistent ratings are”.
* For Supported and Refuted, sentences forming necessary evidence for judgement are recorded
* Best accuracy on labeling a claim accompanied by correct evidence = 31.87%, ignore evidence = 50.91%
* Datasets used for verification and fact-checking are limited to a few hundred claims
* Guidelines to ensure annotation consistency resulted in 95.42% precision, 72.36% recall in evidence retrieval
* Pipeline approach
  + Given claim, identify relevant documents (Claim generation)
  + Select sentences forming the evidence from documents
  + Classify claims w.r.t evidence (Claim labeling)
* Fake News Challenge (Related work)
  + Model verification as stance classification
    - Given claim and article, predict whether article supports, refutes, observes (neutral) or irrelevant to claim
    - Dataset, curated and labelled in context of Emergent Project, classified claim w.r.t article headline instead of whole article
    - Systems provided sources to verify against instead of having to retrieve them
* Claim Generation
  + Extracting info from Wikipedia and generating claims
  + Annotators given random sample of sentences to generate claims from
    - If only source sentence given, claims will be simplifications and paraphrases and thus trivially verifiable
    - If world knowledge freely incorporated, hard to verify claims by Wikipedia alone
    - Given a dictionary - list of terms hyperlinked in original sentence and first sentence from corresponding Wiki page
  + Annotators generate mutations of claims whether supported by Wiki or verified against
    - 6 types: paraphrasing, negation, substitution of an entity/relation with similar/dissimilar one, making claim more general/specific
  + Resulted in claims (extracted mutated) with mean length of 9.4 tokens
* Claim Labeling
  + Classify whether claim supported or refuted by Wikipedia and select evidence, or decide not enough info to make decision
  + Annotators given all sentences from introductory section of page for main entity of claim and every linked entity in those sentences as default source of evidence
  + Annotators record sentences necessary to justify classification decisions
  + Annotators allowed to add arbitrary Wiki page and system adds introductory section as additional sentences that could be selected as evidence
  + NotEnoughInfo label used if claim cannot be supported/refuted by any amount of Wiki info (either because too general/specific)
* Annotation team consists of 50 members, 25 involved only in first task
* Due to complexity of claim labeling, 3 types of data validation used: 5-way inter-annotator agreement, agreement against super-annotators, and manual validation by authors
  + Super-Annotators: expert annotators with no suggested time restriction to annotate randomly selected 1% of data
  + Super-annotators search over whole Wiki for every possible sentence to use as evidence
  + Comparison of regular annotations against this set of evidence is precision/recall 95.42% and 72.36% respectively.
* During data validation, 1.01% of claims skipped, 2.11% contained typos, 6.63% of generated claims flagged as too vague/ambiguous and excluded
* Randomly selected 4% (n=7506) of unskipped claims to be annotated by 5 annotators. Fleiss k score = 0.6841.
* Final quality control step - Chose 227 examples to annotate for accuracy of labels and evidence provided
  + 91.2% examples annotated correctly
  + 3% mistakes in claim generation not flagged during labeling
  + Similar number of claims did not meet guidelines during manual error analysis of baseline system
* When compared against super-annotators, all except two annotators achieved >90% precision and all but 9 achieved recall >70% in evidence retrieval
  + Low-recall cases where super-annotator added >30 sentences as evidence
* During validation by authors, most examples annotated incorrectly where label was correct but evidence selected was not sufficient (only 4 out of 227 examples labeled incorrectly according to guidelines)
  + Tried to resolve issue by asking annotators to err on side of caution
  + “Shakira is Canadian” could be refuted by “Shakira is a Columbian singer…” but advocated that unless more explicit information provided “She was denied Canadian citizenship”, claim should be labeled NotEnoughInfo, since dual citizenships are permitted and annotators’ world knowledge should not be factored in
* Baseline System Description - Simple pipelined system comprising 3 components:
  + Document retrieval
  + Sentence-level evidence selection
  + Textual entailment
  + Each component evaluated in isolation through oracle evaluations on dev set
* Document Retrieval
  + Document retrieval component from DrQA system, returns k nearest documents for query using cosine similarity between binned unigram and bigram Term Frequency-Inverse Document Frequency (TF-IDF) vectors
    - TF-IDF: Numerical statistic intended to reflect how important a word is to a document in a collection or corpus.
    - Often used as weighting factor in searches of information retrieval, text mining and user modeling.
    - Tf-idf value increases proportionally to no. of times word appears in the document and offset by no. of documents in corpus that contain the word, adjusting for fact that some words appear more frequently in general
    - 83% of text-based recommender systems in digital libraries use tf-idf
* Sentence Selection
  + Ranks sentences by TF-IDF similarity to claim
  + Sort most-similar sentences and tune cut-off using validation accuracy on development set
  + Evaluate both DrQA and simple unigram TF-IDF to rank sentences for selection
  + Evaluate impact of sentence selection on RTE module by predicting entailment given original documents w/o sentence selection
    - Textual entailment (TE) - directional relation between text fragments
    - Relation holds true when truth of one text fragment follows from another text
    - In TE framework, entailing and entailed texts are text(t) and hypothesis(h)
    - “t entails h” (t =>h) if a human reading would infer h most likely true
    - Directional because even if “t entails h”, reverse “h entails t” is much less certain
* Recognizing Textual Entailment
  + Compare two models:
    - Simple well-performing baseline - Riedel et al. (2017) Fake News Challenge
      * Multi-layer perceptron (MLP) with single hidden layer, using term frequencies and tf-idf cosine similarity between claim and evidence as features
    - State-of-the-art in RTE - Parikh et al. (2016)
      * Decomposable attention (DA) model between claim and evidence passage
      * Highest scoring system for Stanford Natural Language Inference task
  + RTE must classify claim as NotEnoughInfo when evidence retrieved is not relevant or informative but instances labeled NotEnoughInfo have no evidence annotated, thus cannot train RTE
    - Simulate training instances for NotEnoughInfo using 2 methods:
      * Sampling sentence from nearest page (NearestP) to claim as evidence using document retrieval
      * Sampling sentence from Wiki uniformly at random (RandomS)