Motivation

- Accurately extracting information from text documents is essential for natural language technologies.

- How can we verify if the information is correct by checking it against encyclopedic articles?

This Work

- We introduce a new dataset containing 185,000 true and false facts written by human annotators.

- For each claim, we:
  - add evidence from multiple Wikipedia pages at a sentence level
  - label supported/refuted/not enough info given the evidence.

- Both evidence and label must be correct for scoring. This leads us towards building accountable systems, where a justification/explanation of the verdict is provided.

Dataset Construction

- Claim generation: sentence sampled from intro sections of 50,000 most popular Wikipedia pages. Annotator writes simple sentences for each fact in the original sentence.

- World knowledge can be introduced in controlled manner from a dictionary (using hyperlinked pages on Wikipedia)

- Claim Mutation: for each claim, annotator makes 6 modifications akin to relations in Natural Logic Inference (negation, generalization, specialization, substitution etc.)

- Claim Labelling: different annotator selects a set of sentences that completely support or refute a given claim. Evidence can be combined from multiple pages.

Quality Assurance and Human Evaluation

- Information Retrieval: How annotators with time constraints against super-annotators with no time restrictions? Precision: 95.42%. Recall: 72.36%

- Inference: Are the annotators reaching the same verdict with the evidence they find? We sampled 4% of claims and compute 5-way IAA; Kappa: 0.6841 (n=7506)

- Human Evaluation: We (authors) re-annotated 227 claims, found 91.2% annotated correctly.

- Lessons Learned: Hard to remove annotator’s world knowledge. Hard to come up with ‘universal’ definitions.

Baseline Evaluation

- Evidence Retrieval - DrQA (Chen et al., 2017)
  Trade-off number of documents/sentences (recall) against pipeline RTE accuracy. Upper-bound Score: 62.8%

- Recognizing Textual Entailment
  - Multilayer Perceptron (MLP) (Riedel et al., 2017)
  - Decomposable Attention (DA) (Parikh et al., 2016)

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
<th>FEVER Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP</td>
<td>41.86</td>
<td>19.04</td>
</tr>
<tr>
<td>DA</td>
<td>52.09</td>
<td>32.57</td>
</tr>
</tbody>
</table>

- Future Areas to Explore:
  Multi-sentence natural language inference
  (baseline model concatenates the sentence strings)

  Trade-off between volume of evidence (Recall) and the accuracy of the downstream inference component

  Negative sampling strategy for training textual entailment classifier has a substantial influence on accuracy in real-world setting