Towards Automatic Linkage of Analyst’s Claims with Associated Evidence from Screenshots

Paul Jones, Dakota Medd, Sreekanth Ramakrishnan, Rajat Shah, Joann Keyton, Nagiza F. Samatova, CSC & CHASS, NC State University

Volume and variety of data is continuing to outpace automated methods that help people to analyze it. As a result, knowledge workers falling victim to cognitive biases, particularly confirmation bias.

Need to build computer systems to understand the sensemaking process of analysts and pro-actively help them achieve their goals. This is the purpose of instrumentation but app-specific code is needed to extract content from applications, which becomes complex and brittle; screenshot analysis provides an alternative approach.

Initial research question: can we automatically associate claims with evidence used to support them?

Core assumption: workers are analyzing new information under time pressure; they are not making claims based on prior knowledge of a particular topic, and they have minimal time to contemplate and refactor new knowledge that they learn. Hence we assume that claims and associated evidence will be found close together in time and in space on the computer screen.

Evaluation Results

- Extracted putative claim and evidence sentences and matched using 5 different sentence similarity approaches.
- For the most promising ones, extracted top-100 for human evaluation:
  - 5 raters, each providing a score for the quality of claim/evidence associations using 5-point Likert scale.
  - Compared human to computer scores - pearson correlation of 0.66.
  - Kappa analysis of inter-rater variation showed good agreement for scores 1, 2 and 5.
  - Rater scores:
    - 1/2 - Red
    - 3 - Gray
    - 4/5 - Blue
- Some high scoring pairs:

Future Work and Potential Mission Impact

- Investigate argumentation structure approaches for better claim/evidence detection.
- Investigate more sentence similarity approaches - possibly doc2vec and fasttext.
- Extract more context for evidence (maybe surrounding sentences).
- Attempt multi-source aggregation of evidence -> closer to understanding interpretations.
- Investigate Convolutional Neural Nets for object detection and semantic analysis.
- Key long-term research question: can we build prototypes for analysts that perform automated provenance gathering and maybe automated fact-checking / cross-reference / recommendation capabilities?

Motivation

- Screenshot corpus collected during 2014 LAS/CHASS instrumentation experiment:
  - 150 participants in 54 groups performed a controlled analysis task:
    - ‘Which 3rd party candidate stands the highest chance of winning in the 2016 US Presidential Election?’
    - Users wrote their report in TextEdit app, browsing for evidence in Safari.
  - 121,000 screenshots captured by OSXInstrumenter
    - Periodically every 10 seconds, and on mouse-click and Enter key-pressed events.
  - Key steps in algorithms, and approaches used:
    - Active window detection and boundary detection - template matching from OpenCV.
    - Text extraction/OCR and segmentation - Tesseract and new sentence-length filtering algorithm.
    - Sentence similarity methods - Char sequence (Difflib), BLEU, Word matching (Jaccard), TF/IDF, WordNet.
  - Two algorithm variants evaluated:
    - Stateless (claim and evidence are both in same screenshot e.g. example below)
    - Stateful (evidence can occur in any screenshot prior to the claim being written in TextEdit).

<table>
<thead>
<tr>
<th>Concept</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotation</td>
<td>A manual annotation step that involves identifying and marking evidence in the screenshot.</td>
</tr>
<tr>
<td>Matching</td>
<td>The process of comparing and matching sentences to identify relationships.</td>
</tr>
<tr>
<td>Score Calculation</td>
<td>The method used to assign scores to claim-evidence pairs.</td>
</tr>
</tbody>
</table>

Some problematic cases:

- A: maximized window
- B: wrong active window
- C: multiple TextEdit
- D: highlights and spell check make OCR harder

<table>
<thead>
<tr>
<th>Score Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Poor - no association between sentence</td>
<td>0</td>
</tr>
<tr>
<td>2: Fair - some association but does not match claim.</td>
<td></td>
</tr>
<tr>
<td>3: Good - associated sentences and weak claim evidence.</td>
<td></td>
</tr>
<tr>
<td>4: Very good - looks likely that evidence matches claim.</td>
<td></td>
</tr>
<tr>
<td>5: Excellent - evidence clearly matches claim.</td>
<td></td>
</tr>
</tbody>
</table>