Evaluating an automatic data extraction tool for evidence synthesis through real-life case studies

Linh Hoang, Linh Cao, Yingjun Guan, Yi-Yun Cheng, Jodi Schneider

School of Information Sciences, University of Illinois at Urbana-Champaign

School of Information Sciences
The iSchool at Illinois

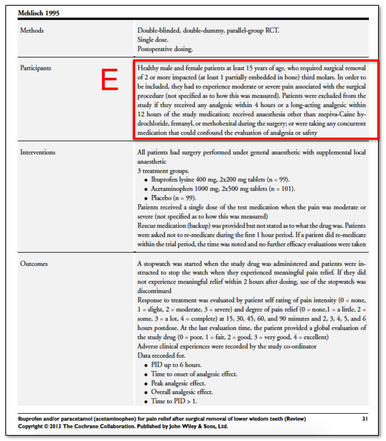
Motivation

- Evidence synthesis is a practice that collects all the available resources of information in order to summarize and interpret existing knowledge.
- In healthcare, evidence synthesis helps to understand how medical knowledge from clinical studies could be transformed into new treatments.
- Our long-term goal is to identify ways to make evidence synthesis fast and effective – with less time and human labor.
- Towards that end, we are evaluating existing computer support tools for key evidence synthesis tasks such as data extraction.

Case Study Design

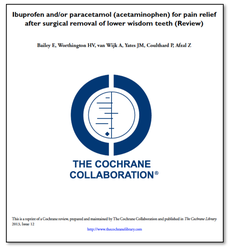
6 included clinical papers

Data Extraction

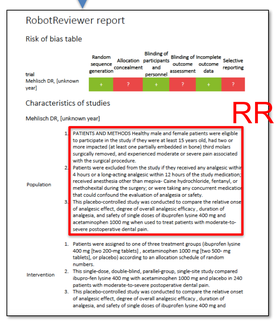


Expert Extraction
(for each included paper)

Synthesized

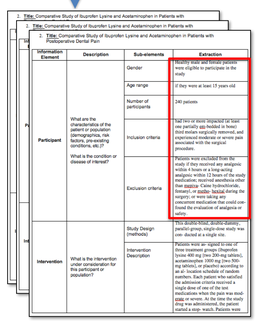


A Cochrane Systematic Review about oral pain relief [1].



RR

RobotReviewer Extraction
(for each included paper)



N_1

3 Novices' Manual Extraction
(for each included paper)

- We are evaluating **RobotReviewer** [2], a tool to help systematic reviewers extract 3 types of data (Population, Intervention, Outcomes) from the full-text of clinical research papers.
- We use the Expert Extraction as a gold standard. We compare its similarity to:
  (1) RobotReviewer Extraction
  (2) Novice_1 Manual Extraction
  (3) Novice_2 Manual Extraction
  (4) Novice_3 Manual Extraction

Evaluate Similarity

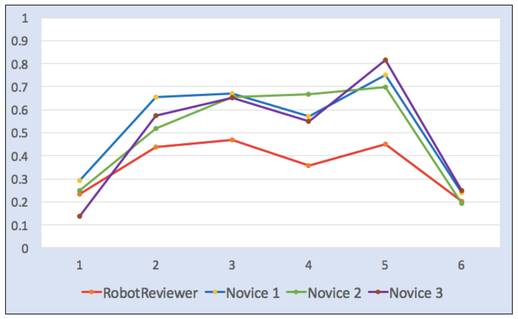
We measure the similarity with the gold standard expert extraction (summary E) using F-measure of **ROUGE-2** [3].

Example of ROUGE-2 (bigram):

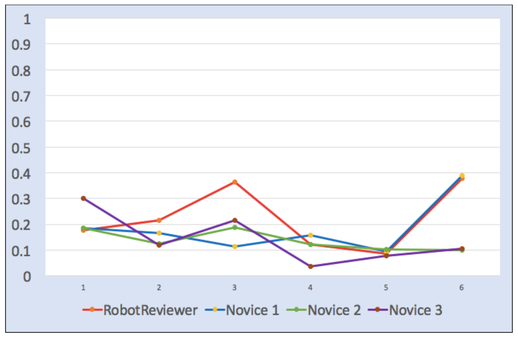
E bigrams: Healthy male and female patients at …
male and female patients …
RR bigrams: Healthy male and female patients were …
male and female patients …
N_1 bigrams: Healthy male and female patients at …
male and female patients …

Overlap bigrams (E, RR): Healthy male, male and, and female, female patients
Overlap bigrams (E, N_1): Healthy male, male and, and female, female patients, patients at

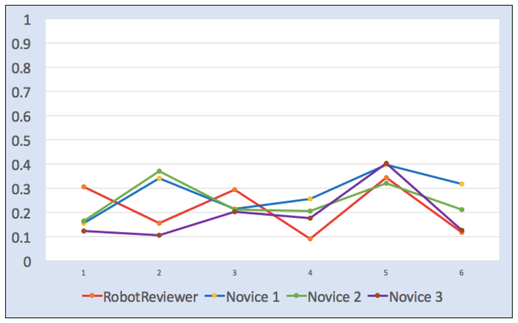
Evaluation Results



Population (F-measure ROUGE-2)
Avg. Mean = 0.47 ; Avg. SD = 0.205; Avg. CV = 43.61



Intervention (F-measure ROUGE-2)
Avg. Mean = 0.172 ; Avg. SD = 0.096; Avg. CV = 56.35



Outcomes (F-measure ROUGE-2)
Avg. Mean = 0.234 ; Avg. SD = 0.097, Avg. CV = 41.61

Discussion

- **Overall RobotReviewer Performance:**
  - The average score of all RR extractions is 0.2-0.3. Performance for Population has the highest average mean.
  - Even though the scores achieved by RR for Intervention & Outcomes are low, they are consistent (with low variation) across the extractions.

- **Population:**
  - Population information (e.g. participant gender, age range) is straightforward. It normally appears at the beginning of full-text content, making it easy to spot.
  - It seems to be the easiest information to extract. Achieves the highest agreement (up to 0.8 for Novice 3 on paper 5).
  - Scores are consistent between extractions (low coefficient of variation - CV), suggesting high agreement between extractors.

- **Intervention:**
  - There is no absolute definition of what should be included in Intervention. Intervention information also can be found in different parts of the full-text content, making it long and varied.
  - It seems to be the hardest information to extract. The highest agreement is only 0.35 (for RobotReviewer on paper 3).
  - Scores vary widely between extractions (high CV), indicating low agreement between extractors.

- **Outcome:**
  - It seems to be hard to extract. The highest agreement achieved is only 0.4 (for Novices 1 & 3 on paper 5).
  - Scores depend on the paper: some vary widely between extractors (with high SD such as for paper 2, paper 6) while others are consistent (with low SD such as for papers 3 & 5).

Ongoing Work

In related work, with expert reviewers at SUNY Buffalo School of Dental Medicine, we are checking how well RobotReviewer would work for an ongoing review about how oral health and systemic health are related.

Future Work

- Analyze inter-annotator agreement of novices.
- Test another evaluation metric, ROUGE-L, which compares two summaries based on the longest common sequences of words.
- Deepen error analysis in order to understand differences between RobotReviewer and novice extraction accuracy.

References

1. Bailey E, Worthington HV, Van Wijk A, Yates JM, Coulthard P, Afzal Z. Ibuprofen and/or paracetamol (acetaminophen) for pain relief after surgical removal of lower wisdom teeth. *The Cochrane Library* 2013.
2. Wallace BC, Kuiper J, Sharma A, Zhu, MB, & Marshall IJ. Extracting PICO Sentences from Clinical Trial Reports using Supervised Distant Supervision. *Journal of Machine Learning Research 17* (2016) 1-25.
3. Lin CY. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop* 2004.

This work was partially funded by the NIH/NLM under R01LM010817