Review Analysis Interview:

Subject12

**7/17/2019 11:00am**

**Interviewee demographics**

*Female*

*Ph.D. in computer science*

*0.5 year working on reviews*

*6 years of data science experience (mostly structured data)*

*Research Scientist at Company2*

# *Reports to other researchers*

# Summary

**Q: Walk us through a recent review analysis task.**

**The data**

*Got hotel reviews, and extractions from [attribute extraction module] (output of another researcher’s work).*

*JSON files, ~millions of reviews which she samples from.*

**So data is clean?**

*No. Extractions have a lot of noise.*

*Interviewee tries to filter out bad extractions by training a small binary classifier-- but collecting labels for this is time-consuming.*

*A major pain point is collecting and auditing crowd worker data. (Even with 3-5 labels per row and testing questions, the labels are often incorrect even for simple tasks.)*

*The interviewee doesn’t share this data because the other researcher uses different methods to evaluate extractions.*

**Goals**

*Analyzing these extractions for relationships within the same review, or the same entity.*

*Creating structures that can support interactive and explainable summarizations from [attribute extraction module] results.*

*Ex: Review says “The room is noisy because it’s right next to the main street, but good location and you can easily walk to the subway.”*

→ The summary should reflect a hierarchy:

* *noisy room*
  + *Next to the main street*
* *central location*
  + *Close to the subway*

*Different granularities of summaries.*

*Later goals:*

*Aggregate the structure over multiple reviews. Might also consider neighborhood summaries at a later stage.*

*This structure might be used to support subjective querying ex: querying children nodes but not parent nodes.*

**Tools:**

*Dumps JSON extractions file into Postgres to leverage indexing by extractions.*

*Custom web app to debug graph model (visualized with D3) of reviews. There is a button to log the index of an interesting example on the server-side.*

*Using D3 to create visualizations is another pain point -- it’s time interviewee could spend on the research problem.*

**Data structures**

*Review graph (or tree) representation*

→ Nodes are extractions

→ Edges are causal relationships

*Still deciding if causal relationships should be restricted to trees.*

*The graph is considered to be noisy evidence, interviewee wants to create heuristics for pruning the graph in order to create a clear hierarchy.*

**Modeling**

*Picks reviews to examine by random sampling*.

Interviewer note: picking examples strategically by the similarity of extractions might be useful

*Uses figure8 to collect labels for randomly chose reviews, uses these to train the edge causality classifier. Interviewee looks for patterns in the errors by eye.*

*Trying to define a good objective function for pruning the graph for the examples we’re looking at. Focus is less on improving the edge classifier than on creating a good pruning heuristic.*

*Eventually, may train this end-to-end.*

*Like others have said, interviewee feels training models is not the bottleneck.*

**Collaboration**

*Dumps interesting things in google doc (graph screenshot)*

*Frequent meetings*

**Bottlenecks**

*Creating labeling tasks, and auditing the results*

*Spends a lot of time hacking custom visualization tools for debugging purposes for the trained model.*

*Scalability is not a bottleneck.*

**Features wish-list**

*Visualization “market place” to support many data structures (especially graphs). Graph visualization tools are limited (unsolved problem), especially for large graphs like hotel sense.*