Review Analysis Interview:

Subject13

**7/18/2019 11:00am**

**Demographics**

**Title** Research Scientist

**Company** Company 2

**Education** Phd

**Experience in data analysis** 2-3 years (mostly structured data)

**Experience in review analysis** 1 year

**Gender** M

# Summary

**Walk us through a recent review analysis task**

*Goal: Extend an attribute extraction pipeline, which was originally designed for hotel and restaurant reviews, to other domains (mechanics, beauty salons, etc.)*

*Extraction examples: aspects (e.g., cleanliness) and modifiers (e.g., “very”) (together called opinions)*

*Pipeline: (1) Data prep/cleaning → (2) Labeling → (3) Modeling*

*Data cleaning happens both before and after data labeling.*

*“Easy” opinions (like “clean room”) are better extracted by rules, so one challenge is finding reviews which have non-”easy” opinions.*

*Before:*

*Filtering*

*Rule-based removal of unsuitable reviews (ex: <3 non-stopword tokens, not English)*

*After:*

*Manually inspect the labeling quality*

*Train a simple (or complex) model on the complete labeled dataset and then check the mispredictions on the training data. These are suspect for cleaning or fixing.*

*Training data quality is critical to the performance of the models. 2% change in training data can lead to 10% improvement in the model.*

*Stops working on improving the quality of labeled data when spent enough time (qualitative--diminishing returns principle). Clearly, though can improve if continued working on it but sometimes it isn’t worth continuing.*

Similar to earlier subjects, (1) interviewee cares much more about ensuring the data quality than model accuracy, (2) debugging is needed at the data curation stage as well as the modeling stage. Again, debugging at the data curation level is, according to the interviewee, more pressing issue than the debugging of the models or their performance. It is partly due to the subjective nature of the task, where what is an “opinion” needs to be qualified by humans.

**What is the goal of your analysis?**

*Extract aspects and their modifiers*

**What kinds of data sources and formats do you use?**

*JSON*

**How do you get them?**

*Public (Yelp, etc.) and private sources (TY)*

**Is scalability a problem?**

*Scalability is an issue when deploying the model and incorporating the unlabeled data (in a semi-supervised setting). It isn’t really a problem earlier in the pipeline b/c labeled set is often much smaller than the original dataset and more isn’t needed and/or expensive*

**How do you iterate on the model?**

*Data augmentation: iterates on different operators and methods of combining them. It might be helpful to look at the results of the data augmentation. Some of the errors are intuitive. For text, there is no perfect operator (usually depends on context).*

*Ex: MixUp data augmentation.*

*Multi-task learning: which tasks to combine?*

*Semi-supervised learning: (like MixMatch) can set parameters for how much you rely on un-labeled data. Perform error analysis on this set of parameters.*

**What tools and languages do you use?**

*PyTorch (used to use Tensorflow), SpaCy, Figure 8, Colab, Jupyter Notebook, Python, BERT (huggingface implementation), Wor2Vec, MixMatch (compare with MixUp), PPT slides, Google Spreadsheets, Jupyter Notebook for visualization*

**What are the downstream applications?**

*Improving experiential search engine*

*Summarization*

*Review featurization*

**What are the bottlenecks you run into or things you spend the most time on? (Are they different?)**

Observe the overlap between the listed bottlenecks and time-consuming aspects

*Bottlenecks:*

*Data labeling --- expensive and not reliable*

*Hard to develop a set of rules to clean data*

*Turnaround time with model training --- too slow (smaller)*

*Time consuming:*

*Data collection*

*Labeling task design*

*Ex: what is an “opinion” in a review?*

*If you can collect training more efficiently and in a more scalable manner, it is relatively easier and faster to train and improve the models*

**What features/tools do you wish you had?**

*Anything that makes it easier to improve the quality of training data*

*Make the labeling task design easier and connect with initial exploratory analysis. Help to figure out what exactly needs to be extracted.*