Review Analysis Interview Subject10

**Demographics**

**Title** Data Analyst in NLP

**Company** Company 4

**Education** PhD in Computer Science

**Experience in data analysis** 6 years (mostly text)

**Experience in review analysis** 1 year

**Gender** F

## Summary

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

Reviews for Hotel industry in Portuguese and Spanish

Purpose: expand the database, find info about entities in DB for Portuguese

* + Spanish and Portuguse KBs are less developed than English, German, French

Attributes are predefined. Use (shallow) grammar rules for extracting opinions.

Predefined schema is 1-to-1 with other language DBs

Works only with review text and schema (no metadata)

**Data preparation**

Reviews are pretty clean (as compared to past projects)

**Task process**

Labeling

* Lexical expansion
  + - Interested in more recent reviews
    - Many shallow measures to estimate level of matches between entities in the reviews and the schema
      * Matching objects to attributes
    - No ground truth
    - Separate labeling tasks
      * Estimating count of “spurious” matches
      * Her team wrote scripts to test the grammar rules
        + Produced many years ago, need updating
        + Language independent
    - Goal is to get to the match level of the English system
    - Testing for precision, not recall
    - Baseline is a statistically based system
    - Portuguese version is in a very early stage of development
    - Spanish is slightly better, still needs a lot of work
    - She does the labeling herself for Portuguese
      * Labels categories of objects
      * Needs to label different inflections and forms of words manually
        + Balances exhaustive labeling with frequency
    - Ruleset is sequential, rules can clash with each other and this introduces difficulty

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

* Plaintext review text (without metadata)
* Mixed corpora of public domain and proprietary data
* Data is collected by a crawling team at the company which obtains data and gives it to her

**Do you often need to look at the raw data points?**

Yes

**Is scalability a problem**

* Yes, cannot use entire data set, but ideally would like to
* Size of Portuguese dataset: nearly 8 million

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

* Python, NLTK, SpaCy, PosGres/SQL, text editors

**How do you collaborate with others**

4 person team for semantic analysis

Just her for Portuguese

**How do you collaborate with others on your work?**

Mostly just communication on Slack, no data sharing

Sharing basic summary statistics

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

* Disambiguation for extracted entities
* Testing and refining new grammatical rules

**Feature requests**

Scripts to search based on lexical features

Wants a visible ontology

Maybe also a dictionary and/or thesaurus?

Less strict ontology

**Other comments**

Can’t use ML due to company prerogative