Review Analysis Interview Subject07

**Demographics**

**Title** Research Scientist/ Engineer and Programmer

**Company** Company1

**Education** Masters in Computational Linguistics

**Experience in data analysis** 20 years (mostly japanese text and some multi-lingual)

**Experience in review analysis** 18 years. Developed text analytics workbench for analyzing sentiments, aspects, targets, syntactic relations.

**Gender** Male

## Summary

**Project goals**: create sentence embeddings to enable subjective search from natural sentence queries

**Data**: About 3 million proprietary reviews. Accessed via BigQuery.

Uses entire corpus to fine-tune BERT embeddings and conduct similarity-based search.

Cleaning

* Extract sentimental dimensions of reviews in order to filter out negative sentences.

(the company provides this as a service for hotel owners, so they don’t want to return negative results)

* Construct positive sentence embedding index.

**Tools:**

BERT embeddings

high dimensional matrix indexing (AISS, facebook)

Python3.7

gensim (for sentiment analysis. Used alongside supervised methods.)

Developing new versions of ginza

Sudachi

**Evaluating the quality of embeddings**:

The target metric is search relevance.

Currently model is unsupervised, but they will compare this to supervised ranking models they tried previously

Beginning stages of project → not quite sure how to evaluate.

**Where do you spend most of your time?**

* Constructing an on-site GPU environment (because scalability is a big factor). NVidia GPU drivers, CUDA libraries setup is very difficult. Once this is done, they want to focus on developing BERT models (where trying a big model is very time-consuming)
* Also, constructed some human-annotated review corpus using CrowdWorks (generally high-quality annotations)

3 corpora:

* Sentimental annotation. Find positive and negative words and their targets in a review.
* Taglining corpus: if you find some positive expression, summarize them to a tagline. (Review summarization.)
* Review/response alignment corpus. In jalan corpus, 80-90% of reviews have responses from the hotel. They aim to find some alignment between review sentences and reply sentences.
  + Annotator makes sentence alignment annotation (ex: 2nd sentence of reply is related to 5th sentence of review.) (Like an entailment task.)
  + Use it to train sentence alignment model with BERT. This is used as a co-training task to generate better sentence embeddings.

**How do you present your results and collaborate?**

Collaborates with an intern who comes in 2 days/week. She’s developing the sentiment analysis algorithm.

Share code and models via github, google cloud (for large models like BERT)

Future plans: Integrate to subjective search functionality. Publish method and library to a company repository, maybe open source.

**Features wish-list:**

Wants to use tensorboard, or anything to help understand/visualize how BERT works.

Suggestion: ElasticSearch 7.3 has high dimensional vector search index. That functionality for evaluating BERT embeddings can be useful.

Also want to visualize the pipeline. (It’s difficult to describe the entire pipeline, through high dimensional search indexing). Docker is useful but complicated to set up.