## Natural Narrative Generation from Structured Data

*Dileep Pasumarthi, Gowri Shankar Ramanan, & Daljeet Virdi*

**Abstract:**  We are creating a program to create narrative summaries from structured data, focused on time series data. Given a sample dataset, we are able to identify interesting trends and develop a narration about those trends. This is useful for telling a story about the dataset automatically. Potential applications include auto-creating news articles, weather reports, industry reports, and business insights. The field we are working on is natural language generation.

**Introduction:** The problem we are trying to solve is automating insights and intelligence from structured data. This is important in order to create engaging reports about data, an otherwise laborious process. Our approach toward doing this with natural language generation will make automatic an otherwise highly human intensive type of work. Lowering the cost to generate these reports will allow for much more of them to be created and

explored.

**Related Work:** There are a few products that have tried similar kinds of work. These include, Narrative Science, Yseop, Arria, AX Semantics, and Automated Insights. However, all of them are rudimentary and verbose. They offer basic summarization showing information like: min, max, average. They do not offer insights from anomaly, causal, or trend detection. These include:

* Distribution
  + Outlier detection
  + Change detection
  + Novelty detection
* Comparisons
* Relationships
  + Regression analysis
  + Necessary condition analysis
* Correlations

**Problem Definition:** Given time series data, compute insights and intelligence and return them as a coherent natural language narration. The input is time series data. The output is a paragraph or two of text. For the purpose of this project, we focus on weather data, although the use cases are much broader and we want to make a generalizable tool that is dataset agnostic. There are use cases for this technology in the retail and restaurant industry, weather correlates to in store traffic, as well as in the transportation industry, as weather correlates with airline and train reliability and departure times.

We want to try answering as many of the following questions as we can:

|  |  |  |  |
| --- | --- | --- | --- |
| **Task** | **General**  **Description** | **Pro Forma**  **Abstract** | **Examples** |
| **Retrieve Value** | Given a set of specific cases, find attributes of those cases. | What are the values of attributes {X, Y, Z, ...} in the data cases {A, B, C, ...}? | *- What is the weather like on Oct 11, 2019?*  *- What is the wind direction on Oct 8, 2019?* |
| **Filter** | Given some concrete conditions on attribute values, find data cases satisfying those conditions. | Which data cases satisfy conditions {A, B, C...}? | *- What weeks of the year are the hottest?*  *- What days of the year are the windiest?* |
| **Compute Derived Value** | Given a set of data cases, compute an aggregate numeric representation of those data cases. | What is the value of aggregation function F over a given set S of data cases? | *- What is the average weather for the month of October?*  *- How much variation is there in the weather in a given zip code?* |
| **Find Extremum** | Find data cases possessing an extreme value of an attribute over its range within the data set. | What are the top/bottom N data cases with respect to attribute A? | *- What is the day with the hottest weather?*  *- What month of the year is the coldest?* |
| **Sort** | Given a set of data cases, rank them according to some ordinal metric. | What is the sorted order of a set S of data cases according to their value of attribute A? | *- Order the months by year.* |
| **Determine Range** | Given a set of data cases and an attribute of interest, find the span of values within the set. | What is the range of values of attribute A in a set S of data cases? | *- What is the range of weather for Fall?*  *- What is the range of wind for October?* |
| **Characterize Distribution** | Given a set of data cases and a quantitative attribute of interest, characterize the distribution of that attribute's values over the set. | What is the distribution of values of attribute A in a set S of data cases? | *- What is the distribution of temperature in Winter in Palo Alto?* |
| **Find Anomalies** | Identify any anomalies within a given set of data cases with respect to a given relationship or expectation, e.g. statistical outliers. | Which data cases in a set S of data cases have unexpected/exceptional values? | *- Are there exceptions to the relationship between wind and temperature?*  *- Are there any outliers in temperature?* |
| **Cluster** | Given a set of data cases, find clusters of similar attribute values. | Which data cases in a set S of data cases are similar in value for attributes {X, Y, Z, ...}? | *- Are there groups of zip codes w/ similar temperature and humidity?*  *- Is there a cluster of similar temperature zip codes?* |
| **Correlate** | Given a set of data cases and two attributes, determine useful relationships between the values of those attributes. | What is the correlation between attributes X and Y over a given set S of data cases? | *- Is there a correlation between temperature and wind?*  *- Is there a correlation between humidity of latitude?* |
| [**Contextualization**](https://en.wikipedia.org/wiki/Contextualization_(computer_science)) | Given a set of data cases, find contextual relevance of the data to the users. | Which data cases in a set S of data cases are relevant to the current users' context? | *- Are there groups of zip codes that have similar temperature?* |

**Solution:**

We are initially using three years of weather data for a number of zip codes in the US and narrate interesting facts about it.

Our method works as follows:

1. Clean and sanitize weather data.
   1. [Scrape data from weather underground.](https://www.wunderground.com/history/daily/KPAO/date/2019-10-10?req_city=Palo%20Alto&req_state=CA&req_statename=California&reqdb.zip=94306&reqdb.magic=1&reqdb.wmo=99999)
2. Find interesting insights about the data using:
   1. Anomaly detection
      1. [Nearest neighbors + Local Outlier Factor](https://pdfs.semanticscholar.org/b4c9/3848b0808566b06e6527901fd07a3aa2a2f4.pdf)
      2. Causal discovery
   2. Clustering
   3. Time series data
   4. Structured learning
3. Converting automated insights into narration

**Identified technologies we may employ:**

1. Public weather data
2. A few libraries to find interesting insights from data such as:
   1. [Structured Learning used at Uber](https://github.com/mauro-idsia/blip)
   2. Time Series Data:
      1. [TsFresh](https://tsfresh.readthedocs.io/en/latest)
      2. [Dtaidistance](https://github.com/wannesm/dtaidistance)
   3. Anomaly Detection
      1. [anomatools](https://github.com/Vincent-Vercruyssen/anomatools%5D(https://github.com/Vincent-Vercruyssen/anomatools))
   4. Causal Discovery
      1. [causalDisco](https://github.com/annennenne/causalDisco%5D(https://github.com/annennenne/causalDisco))
   5. Clustering:
      1. [spectral-clustering](https://github.com/pin3da/spectral-clustering%5D(https://github.com/pin3da/spectral-clustering))
3. Convert insights into narration
   1. [simplenlg](https://github.com/simplenlg/simplenlg)

### **Demo:** We plan to show our demo by taking raw weather data from any geography and returning a narration showcasing trends and insights automatically.

### **Conclusion and future work:** With the ubiquity and rapid growth of data, we don’t have enough data analysts to understand and make sense of all of it. Using a tool like that converts structured data to insights and narrating them can auto generate key reports about that data. Imagine a daily or weekly report about the weather or any other sort of structured data without needing a human in the loop. Further, using any of the off the shelf commoditized text to speech algorithms, you could listen to this same narration while driving or on the bus. If this is successful, large portions of a weatherman’s day job completely automated - currently a highly labour intensive process. If we can automate this process, we reduce the cost of analyzing and synthesizing insights from data by many orders of magnitude.

### **References:**

[1] <https://www.bluewolf.com/bluewolf-now/future-retail-weather-insights-qa-ibms-paul-walsh>

[2] <https://www.earthnetworks.com/blog/5-key-insights-from-our-airport-weather-report-view-infographic/>

[3] <https://enterprise.foursquare.com/intersections/article/summer-weather-data-reveals-consumer-insights>