Semantic Concept Discovery Over Event Data

Oktie Hassanzadeh, Shari Trewin, Alfio Gliozzo
IBM Research AI
Use Case: Question Analysis

• High-Level Goal:

  Given a question (set of entities & concepts), find the most relevant entities & concepts needed to generate a high-quality analysis report.

• Example question:
  “What are the consequences of Brexit on London’s financial markets?”

Need to discover:
  – Key topics (e.g., financial markets, economy, Brexit, Brexit Divorce Bill)
  – Key people and organizations involved (e.g., The European Union, decision makers in the EU & UK, people involved in Brexit negotiations)
  – All the related events (e.g., Negotiation meetings, Parliamentary elections within the EU, etc.)
Concept Discovery for Deep Analysis

Goal: Given an analysis question, find the most relevant concepts (topics, entities) needed to make a comprehensive and unbiased analysis

Classic Solution:

**Search** over news articles, social media, past reports, other media (text, photo & videos)

**Identify** relevant concepts mentioned in unstructured sources to establish context

**Perform Analysis** tasks such as forecasting, hypothesis generation, and scenario analysis, given the context

Problem #1: Mostly manual process, and massive amount of information - so the outcome could be **biased & incomplete**

Problem #2: Sources of search do not match the available structured data and the corresponding models used for **analytics**
Event Databases

- Structured data representing "events" as reported on the media

GDELT  https://www.gdeltproject.org/


EventRegistry  http://eventregistry.org/
Event Databases

- Structured data representing "events" as reported on the media

   - **GDELT**
     - Political event databases: “Event” is an action associated with up to two actors

   - **ICEWS**

   - **EventRegistry**
     - Generic event database: “Event” is a collection of articles on the same topic
Event Databases

- Structured data representing "events" as reported on the media

- **GDELT 2.0**
  - 129+ Million Events
  - 157+ Million Articles Annotated (GKG)
  - 437+ Million Mentions (2+ years)

- **ICEWS**
  - 14.9 Million Events (1995 to 2015)

- **EventRegistry**
  - 5+ Million Events
  - 180+ Million Annotated Articles (2+ years)
Concept Discovery for Deep Analysis: Our Solution

Semantic Data Curation Engine:
- GDELT: 129 Million Events, 157 Million Annotations, 437 Million Mentions.
- ECNWS: 3 Million Events, 34+ Million Annotated Articles.
- EventRegistry: 14.9 Million Events.
- FactBase: 6.3 Billion Triples, 488 Million Objects.

Unified Index

Semantic Embeddings Engine

Analysis Question: What is the likelihood of violent protest in Caracas, Venezuela?

Natural Language Question Understanding:
- Topic: Violent Protest
- Location: Caracas, Venezuela

Semantic Concept Discovery:
- Query DeepSim API with Topic#Violent_Protest Location#Caracas_Venezuela

DeepSim Analysis

Key People:
- Jesus Torrealba
- Henrique Capriles
- Nicolas Maduro
- Ewaner Chaves
- David Cristobal

Organizations:
- Venezuela Supreme Court
- United Socialist Party Of Venezuela
- National Electoral Council
- Venezuelan National Assembly

Themes:
- Tax Ethnicity Venezuelan
- Tax Ethnicity Venezuelans
- Tax Political Party Unity Alliance
- Tax From FarewellAgreements Memoir
Question Analysis: Current Prototype

Input could be a question, or a set of entities and topics
Question Analysis: Current Prototype

Input could be a question, or a set of entities and topics
We have ingested 3 major event databases
All are structured (tabular) or semi-structured (JSON) data
Context is identified through lookups in our FactBase API
Question Understanding: Context Details

Cognitive Assistant for the Analyst

Question Analysis

Brazil, Dilma Rousseff, Impeachment

Global Context

DeepSim Analysis

Powered by IBM Socrates & GDELT

DeepSim Context:

Key People

Dilma Rousseff
Eduardo Cunha
Renan Calheiros
Luiz Inacio Lula
Waldir Maranhao
Aecio Neves
Senio Moro

Facts from IBM FactBase

Links
http://www.wikidata.org/entity/Q40722

Occupation: Politician, Economist
Position Held: President Of Brazil
Educated At: Federal University Of Rio Grande Do Sul
Award Received: Bertha Lutz Prize, Order Of Isabella The Catholic
Father: Pedro Rousseff
Follows: Luiz Inacio Lula Da Silva
Country Of Citizenship: Brazil
Member Of Political Party: Workers’ Party
Place Of Birth: Belo Horizonte
Description: President Of Brazil
Label: Dilma Rousseff
Url: http://www.wikidata.org/entity/Q40722
DeepSim Analysis – Using Semantic Similarity Analysis over GDELT GKG

Concepts are mapped from FactBase IDs to GDELT GKG terms

Results using embeddings built over 157+ million GDELT GKG records
Swagger API behind DeepSim Analysis

![Swagger API Image]

Implementation Notes:
Similarity Analysis for events using deep learning.

Response Class
Model: Model Schema
array(string)

Response Content Type: application/json

Parameters:
- **terms**
  - Value: comma-separated list of source terms (field_name=value).
  - Description: query
  - Parameter Type: string
- **targetField**
  - Value: person
  - Description: target field name. Example: name, organization, person, theme, location_fullName, location_countryCode, location_latitudeLongitude
  - Parameter Type: string
- **limit**
  - Value: 20
  - Description: number of terms in the output.
  - Parameter Type: integer
- **model**
  - Value: cgx2c_v1(default)
  - Description: Model.
  - Parameter Type: string

Request URL:
http://oliver.s1.cloud9.ibm.com:7070/rest/api/similarity/deepSimMulti?terms%20(sourceField%23value)=location_fullName

Response Body:
```
{
  "result": [  "person#Bima_Roussel",
              "person#Barbaro_Cume",
              "person#Mersi_Caiheng",
              "person#Zhang_Inscio_Lula",
```


Index-based Analysis (co-occurrence)

Results using queries over our event databases (EventRegistry for this view)
Swagger API behind Index-Based Analysis

**concepts-discovery**
This is the concept discovery API. It provides a set of functions for expanding a query with related people, organizations and themes.

**extract**
```
POST /extract/concepts
This api extracts concepts (person, organization, themes, etc.) from a given text string
```

**query**
```
POST /query/expand
This api uses GKG data to find key players (people and organizations) for a given country, set of seed people, and set of GDELT themes (topics)
Example Input: { "country": "Peru", "players": ["Nicolas Maduro"], "topics": ["IMMIGRATION", "PROTEST"], "candidateRankingMethods": ["similarity", "facetcount"] }
```

Response Class (Status 200)
successful operation

<table>
<thead>
<tr>
<th>Model</th>
<th>Example Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```json
{
   "field": "string",
   "method": "string",
   "candidates": ["string"
```
Evaluation

- Benchmark created using reports written by human experts
  - Human Rights Watch organization & Wikipedia articles on events and people
- We measured the ability of different algorithms to find the concepts mentioned in the original reports
- A combination approach works best in most cases
Conclusion & Future Work

700+ million records
24+ billion values

Structured Event Data

Concept Index

Semantic Embeddings

Concept Extraction

Concept Discovery

Concept Ranking

Visualization, Hypothesis Generation, Causal Reasoning, Forecast Engine

Analyst Input
(Natural Language Query, Document, Image, Video)

Event Extraction Engine

Semantic vectors for every field/value using various (temporal) models

Developed a benchmark & thoroughly evaluated - showing very promising results. Larger size and better quality benchmark is needed.

Exploration of key players, locations, events, and scenarios
We are hiring!

and we very much welcome academic collaborations

Come to our booth & get in touch.

(You may take one of ☝ these home!)
BACKUP
System Architecture

Structured Knowledge Sources
- Wikidata
- DBpedia
- YAGO
- Freebase

Event Databases
- GDELT
- ICEWS
- GKG
- EventRegistry

Ingestion:
- Crawl, Parse, Clean/Filter, Store

Curation:
- Pre-process, Match, Index

Event Knowledge Graph & Concept Discovery APIs

Semantic Term Embeddings Creation

Embeddings Management System
DeepSim (context) Analysis Details – Model Construction & Query Engine

• Step 2: Model Construction & Query Engine

- Word2vec modifications
  - Fixing context window size, rotating window so column order does not affect the outcome
- Embeddings Management System
  - Super fast in-memory approximate nearest neighbor library & API
- Query Engine for DeepSim
  - Prefix-based query for field-specific retrieval (e.g., retrieve ”person”s similar to “location”)
  - AND query over input terms (e.g., retrieve “person” similar to “location X”, “persons Y & Z”
Concept Ranking: Index-Based Method (co-occurrence)

- Index-based Method: Measuring Co-occurrence
  - Formulate a search query using the concepts extracted from the input question
  - Count the concept annotations for every record in the input
  - Return the most frequent annotations of various types (persons, organizations, themes)
  - Use percentage of co-occurrence as the relevance score
Concept Ranking: Deep Similarity Method (context)

- Using word embeddings to capture the semantic similarity of terms (field values)
- Embeddings in NLP: vectors representing the semantic context of each word
  - Similar terms have similar vectors (as per e.g. Cosine similarity between the vectors)
- Method used: modified Skip-gram word2vec model [Mikolov et al., 2013]
  - An efficient, shallow Neural Network model
  - CBOW architecture predicts the current word based on the context
  - Skip-gram predicts surrounding words given the current word
DeepSim (context) Analysis Details – Virtual Document Generation

• Step 1: Virtual Document (Context) Generation
  – A term is a concatenation of a field name (column header) and a field value (cell content)
  – Each record in the input data turns into a context
  – We may need transformations such as data binning / bucketing for numerical fields

<table>
<thead>
<tr>
<th>date</th>
<th>Location_countryCode</th>
<th>Location_fullName</th>
<th>persons</th>
<th>names</th>
<th>themes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017-01-01</td>
<td>AS</td>
<td>Goulburn, New South Wales, Australia</td>
<td>Barnaby Joyce, Alastair Starritt, ...</td>
<td>Water Minister Barnaby, Southern Valley, ...</td>
<td>GENERAL GOVERNMENT, EPU_POLICY_GOVERNMENT, ...</td>
</tr>
</tbody>
</table>

2017-01-01T08:00:00Z tone_selfGroupReferenceDensity#0 tone_tone#-2 tone_activity#2 tone_positiveScore#0 tone_negativeScore#0 person#Barnaby_Joyce person#Alastair_Starritt person#Alastair_Starritt sourceCommonName#stockandland_com_au organization#Deniboota_Landholders_Association organization#Deniboota_Landholders_Association name#Deniboota_Landholders_Association name#Alastair_Starritt name#Deniboota_Landholders_Association name#Murray-Darling_Basin name#Northern_Basin name#Basin_Plan name#Southern_Basin name#Goulburn_Murray name#Goulburn_Valley name#Basin_Plan name#Latrobe_Valley name#Water_Minister_Barnaby_Joyce sourceCollectionIdentifier#WEB theme#GENERAL_GOVERNMENT theme#GENERAL_GOVERNMENT theme#GENERAL_GOVERNMENT theme#EPU_POLICY_GOVERNMENT ... location_latitudeLongitude#149.721--34.7515 location_countryCode#AS location_fullName#Goulburn_New_South_Wales_Australia location_type#WORLDCITY location_featureID#-1576139 locationADM1Code#AS02 locationADM2Code#154641 location_latitudeLongitude#149.721--34.7515 location_countryCode#AS location_fullName#Goulburn_New_South_Wales_Australia location_type#WORLDCITY location_featureID#-1576139 ...
Towards a Generic Pipeline for Semantic Similarity Analysis

- Input Structured, Semi-structured, and Unstructured Data & Knowledge
- Scale-Out Data Curation Platform
  - Mapping, Integration, Discovery, Curation
  - Knowledge Graph
- Deep Similarity Analysis Pipeline
- Decision Analytics Pipeline
- Interactive Discovery
The Need for an Event Extraction Engine

• Shortcomings of existing event data
  – No associated text articles. Only URLs are available.
  – No associated meta-data (GKG or GKG-like data) over historic articles -> we only have GKG data limited to the past two years.
  – Limited definition of "event" in each source. E.g., GDELT & ICEWS are CAMEO coded (and so only cover a specific type of political events) and EventRegistry defines an event as a collection of articles so we do not know the kind of actions and actors in each event.
  – Noise \(\rightarrow\) both random noise and systematic noise (a result of rule-based extraction)
    – EventRegistry
      – encoding issues (tofu characters in labels)
      – inaccurate concept annotations
    – GDELT & ICEWS
      – wrong annotations, missing annotations, basically all the problems a rule-based system could have

• What we need: a comprehensive, accurate, and up-to-date event database with annotations similar to EventRegistry and GDELT GKG, historic coverage similar to ICEWS, event coding similar to GSR