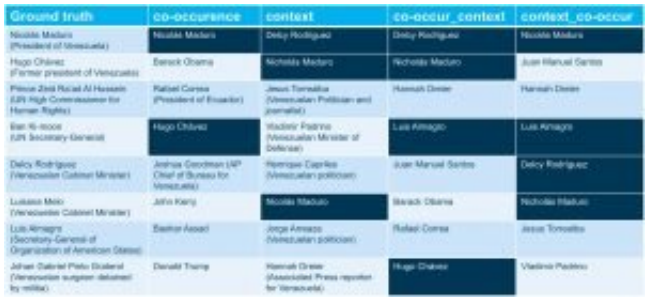
Semantic concept discovery over event databases

17 July 2018, by Oktie Hassanzadeh



Comparison of concept rankings for a Human Rights Watch Report. The 'Ground truth' column shows the eight most frequently mentioned people in the 'Venezuela's Humanitarian Crisis' report, while the other columns shows values returned by various discovery methods. Values that are among the ground truth concepts are indicated by dark boxes . The context method returns values that are all relevant (even if missing from the original article), whereas the co-occurrence method returns many popular but irrelevant concepts (e.g., politicians making general statements on the topic). Credit: IBM

At IBM Research AI, we built an AI-based solution to assist analysts in preparing reports. The paper describing this work recently won the best paper award at the "In-Use" Track of the 2018 Extended Semantic Web Conference (ESWC).

Analysts are often tasked with preparing comprehensive and accurate reports on given topics or high-level questions, which may be used by organizations, enterprises, or government agencies to make informed decisions, reducing the risk associated with their future plans. To prepare such reports, analysts need to identify topics, people, organizations, and events related to the questions. As an example, in order to prepare a report on the consequences of Brexit on London's financial markets, an analyst needs to be aware of the key related topics (e.g., financial markets, economy, Brexit, Brexit Divorce Bill), people and organizations (e.g., The European Union, decision makers in the EU & UK, people involved in Brexit negotiations), and events (e.g., Negotiation meetings, Parliamentary elections within the EU, etc.). An AI-assisted solution can help analysts to prepare complete reports and also avoid bias based on past experience. For example, an analyst could miss an important source of information if it has not been used effectively in the past.

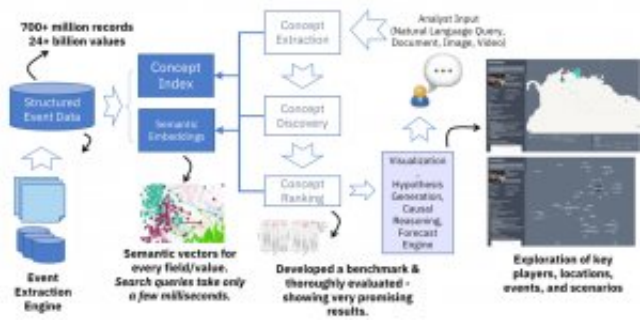
The knowledge induction team at IBM Research AI built the solution using deep learning and structured event data. The team, led by Alfio Gliozzo, also won the prestigious Semantic Web Challenge award last year.

Semantic embeddings from event databases

The key technical novelty of this work is the creation of semantic embeddings out of structured event data. The input to our semantic embeddings engine is a large structured data source (e.g., database tables with millions of rows) and the output is a large collection of vectors with a constant size (e.g., 300) where each vector represents the semantic context of a value in the structured data. The core idea is similar to the popular and widely used idea of word embeddings in natural language processing, but instead of words, we represent values in the structured data. The result is a powerful solution enabling fast and effective semantic search across different fields in the database. A single search query takes only a few milliseconds but retrieves results based on mining hundreds of millions of records and billions of values.

While we experimented with various neural network models for building embeddings, we obtained very promising results using a simple adaptation of the original skip-gram word2vec model. This is an efficient shallow neural network model based on an architecture that predicts the context (surrounding words) given a word in a document. In our work, we are dealing not with text documents but with

structured database records. For this, we no longer need to use a sliding window of a fixed or random size to capture the context. In structured data, the context is defined by all the values in the same row regardless of the column position, since two adjacent columns in a database are as related as any other two columns. The other difference in our settings is the need to capture different fields (or columns) in the database. Our engine needs to enable both general semantic queries (i.e., return any database value related to the given value) and field-specific values (i.e., return values from a given field related to the input value). For this, we assign a type to the vectors built out of each field and build an index that supports type-specific or generic queries.



Credit: IBM

For the work described in our paper, we used three publicly available event databases as input: GDELT, ICEWS, and EventRegistry. Overall, these databases consist of hundreds of millions of records (JSON objects or database rows) and billions of values across various fields (attributes). Using our embeddings engine, each value turns into a vector representing the context in the data.

A simple retrieval query

One can see how well the context is captured by our engine using a simple retrieval query. For example, when querying for value "Hilary Clinton" (misspelled) in field "person" in GDELT GKG, the first hit or most similar vector is "Hilary Clinton" (misspelled) under field "name" and the next most similar vectors are "Hillary Clinton" (correct spelling) under fields "person" and "name". This is due to the very similar context of the misspelled value and the correct spelling, and also the values across the fields "name" and "person". The rest of the hits for the above query include U.S. politicians, particularly those active during the last presidential elections, as well as related organizations, persons with similar job roles in the past, and family members.

Similarity search on combined queries

Of course, our solution is capable of achieving much more than a simple retrieval query. In particular, one can combine these queries to turn a set of values extracted from a natural language query into a vector and perform similarity search. We evaluated the outcome of this approach using a benchmark built from reports written by human experts, and examined the ability of our engine to return the concepts described in the reports using the title of the report as the only input. The results clearly showed the superiority of our semantic embeddings–based concept discovery approach compared with a baseline approach relying only on the co-occurrence of the values.

New applications in concept discovery

A very interesting aspect of our framework is that any value and any field is assigned a vector representing its context, which enables new interesting applications. For example, we embedded latitude and longitude coordinates from events in the databases into the same semantic space of concepts, and worked with the Visual AI Lab led by Mauro Martino to build a visualization framework that highlights related locations on a geographic map given a question in natural language. Another interesting application we are currently investigating is using the retrieved concepts and their semantic embeddings as features for a machine learning model that the analyst needs to build. This can be used in an automated machine learning and data science (AutoML) engine, and support analysts in another important aspect of their jobs. We are planning to integrate this solution in IBM's Scenario Planning Advisor, a decision support system for risk analysts.

More information: Semantic Concept Discovery Over Event Databases. 2018.eswc-conferences.org/paper_182/

This story is republished courtesy of IBM Research. Read the original story here.

Provided by IBM