Semantic Concept Discovery over Event Databases

Oktie Hassanzadeh(✉), Shari Trewin, and Alfio Gliozzo

IBM Research, Yorktown Heights, NY, USA
hassanzadeh@us.ibm.com

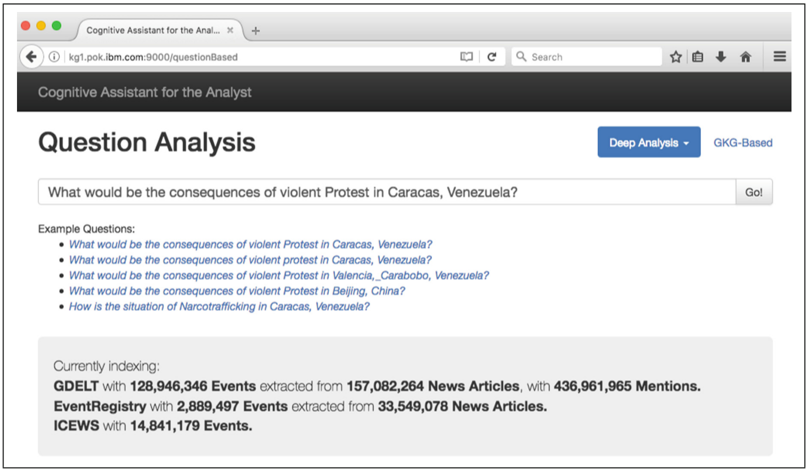
Abstract. In this paper, we study the problem of identifying certain types of concept (e.g., persons, organizations, topics) for a given analysis question with the goal of assisting a human analyst in writing a deep analysis report. We consider a case where we have a large event database describing events and their associated news articles along with meta-data describing various event attributes such as people and organizations involved and the topic of the event. We describe the use of semantic technologies in question understanding and deep analysis of the event database, and show a detailed evaluation of our proposed concept discovery techniques using reports from Human Rights Watch organization and other sources. Our study finds that combining our neural network based semantic term embeddings over structured data with an index-based method can significantly outperform either method alone.

1 Introduction

Analysts are often tasked with preparing a comprehensive, accurate, and unbiased report on a given topic. The first step in preparing such a report is a daunting discovery task that requires researching through a massive amount of information. Information sources can have large volume, variety, varying veracity, and velocity - the common characteristics of the so-called Big Data sources. Many times the analysis requires a deep understanding of various kinds of historical and ongoing *events* that are reported in the media. To enable better analysis of events, there exist several *event databases* containing structured representations of events extracted from news articles. Examples include GDELT [17], ICEWS [11], and EventRegistry [16]. These event databases have been successfully used to perform various kinds of analysis tasks, e.g., forecasting societal events [22]. However, there has been little work on the discovery aspect of the analysis, that results in a gap between the information requirements and the available data, and potentially a biased view of the available information.

In this paper, we present a framework for concept discovery over event databases using semantic technologies. Unlike existing concept discovery solutions that perform discovery over text documents and in isolation from the remaining data analysis tasks [18,28], our goal is providing a unified solution that allows deep understanding of the same data that will be used to perform

© Springer International Publishing AG, part of Springer Nature 2018
A. Gangemi et al. (Eds.): ESWC 2018, LNCS 10843, pp. 288–303, 2018.
https://doi.org/10.1007/978-3-319-93417-4_19


Cognitive Assistant for the Anal...
kg1.pok.ibm.com:9000/questionBased
Search
Deep Analysis ▾
GKG-Based
Go!

Cognitive Assistant for the Analyst

Question Analysis

What would be the consequences of violent Protest in Caracas, Venezuela?

Example Questions:

- *What would be the consequences of violent Protest in Caracas, Venezuela?*
- *What would be the consequences of violent protest in Caracas, Venezuela?*
- *What would be the consequences of violent Protest in Valencia,_Carabobo, Venezuela?*
- *What would be the consequences of violent Protest in Beijing, China?*
- *How is the situation of Narcotrafficking in Caracas, Venezuela?*

Currently indexing:
GDELT with **128,946,346 Events** extracted from **157,082,264 News Articles**, with **436,961,965 Mentions.**
EventRegistry with **2,889,497 Events** extracted from **33,549,078 News Articles.**
ICEWS with **14,841,179 Events.**

Fig. 1. Question analysis UI - main page

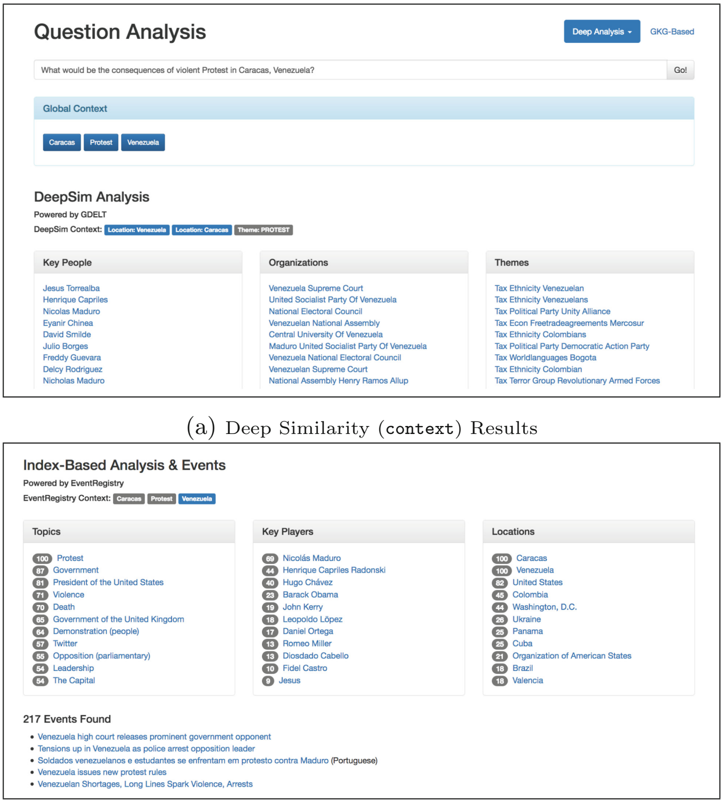
other analysis tasks (e.g., hypothesis generation [27], scenario planning [26], or building models for forecasting [15,22]). Figures 1 and 2 show different views of our system's UI that is built using our concept discovery framework APIs. The analyst can enter a natural language question or a set of concepts, and retrieve collections of relevant concepts identified and ranked using different discovery algorithms described in Sect. 3. Using this system provides a new starting point for an analyst's work. Instead of performing complex search queries and examining pages of results, the analyst reviews the related concepts, exploring what connects them to the analytical question. Unexpected concepts broaden the scope of their thinking, helping to overcome confirmation bias. A key aspect of our framework is the use of semantic technologies. In particular:

- A unified view over multiple event databases and a background RDF knowledge base is achieved through semantic link discovery and annotation.
- Natural language or keyword query understanding is performed through mapping of input terms to the concepts in the background knowledge base.
- Concept discovery and ranking is performed through neural network based semantic term embeddings.

In what follows, we first describe the overall framework and its various components. We then describe the algorithms used for concept discovery and ranking. In Sect. 4, we present the methodology and results of our evaluation using a ground truth built from a large corpus of reports written by human experts.

2 Concept Discovery Framework

Figure 3 shows the architecture of our system. The system takes in as input a set of event databases and RDF knowledge bases and provides as output a set of



Question Analysis

What would be the consequences of violent Protest in Caracas, Venezuela?

Global Context

Caracas Protest Venezuela

DeepSim Analysis

Powered by GDELT

DeepSim Context: Location: Venezuela Location: Caracas Theme: PROTEST

Key People	Organizations	Themes
Jesus Torrealba	Venezuela Supreme Court	Tax Ethnicity Venezuelan
Henrique Capriles	United Socialist Party Of Venezuela	Tax Ethnicity Venezuelans
Nicolas Maduro	National Electoral Council	Tax Political Party Unity Alliance
Eyanir Chinea	Venezuelan National Assembly	Tax Econ Freetradeagreements Mercosur
David Smilde	Central University Of Venezuela	Tax Ethnicity Colombians
Julio Borges	Maduro United Socialist Party Of Venezuela	Tax Political Party Democratic Action Party
Freddy Guevara	Venezuela National Electoral Council	Tax Worldlanguages Bogota
Delcy Rodriguez	Venezuelan Supreme Court	Tax Ethnicity Colombian
Nicholas Maduro	National Assembly Henry Ramos Allup	Tax Terror Group Revolutionary Armed Forces

(a) Deep Similarity (`context`) Results

Index-Based Analysis & Events

Powered by EventRegistry

EventRegistry Context: Caracas Protest Venezuela

Topics		Key Players		Locations	
100	Protest	69	Nicolás Maduro	100	Caracas
87	Government	44	Henrique Capriles Radonski	100	Venezuela
81	President of the United States	40	Hugo Chávez	82	United States
71	Violence	23	Barack Obama	45	Colombia
70	Death	19	John Kerry	44	Washington, D.C.
65	Government of the United Kingdom	18	Leopoldo Löpez	26	Ukraine
64	Demonstration (people)	17	Daniel Ortega	25	Panama
57	Twitter	13	Romeo Miller	25	Cuba
55	Opposition (parliamentary)	13	Diosdado Cabello	21	Organization of American States
54	Leadership	10	Fidel Castro	18	Brazil
54	The Capital	9	Jesus	18	Valencia

217 Events Found

- Venezuela high court releases prominent government opponent
- Tensions up in Venezuela as police arrest opposition leader
- Soldados venezuelanos e estudantes se enfrentam em protesto contra Maduro (Portuguese)
- Venezuela issues new protest rules
- Venezuelan Shortages, Long Lines Spark Violence, Arrests

(b) Index-Based (`co-occur`) Results

Fig. 2. Question analysis UI - concept discovery results

APIs that provide a unified retrieval mechanism over input data and knowledge bases, and an interface to a number of concept discovery algorithms. In what follows, we describe the input sources and each of the components in detail.

2.1 Event Data and Knowledge Sources

Event databases are structured records describing various kinds of societal, political, or economic events. While event extraction from text is a well-studied topic in the NLP literature [12,14] with a dedicated track at the annual Text Analysis Conference (TAC) [6], there are only a few publicly available large-scale event databases. The input of these event databases is a large corpus of news articles that are either gathered from various news sources (e.g., news agencies and other proprietary sources) or crawled from the Web. The output is structured records (i.e., relational data tables) describing various features of the identified events.

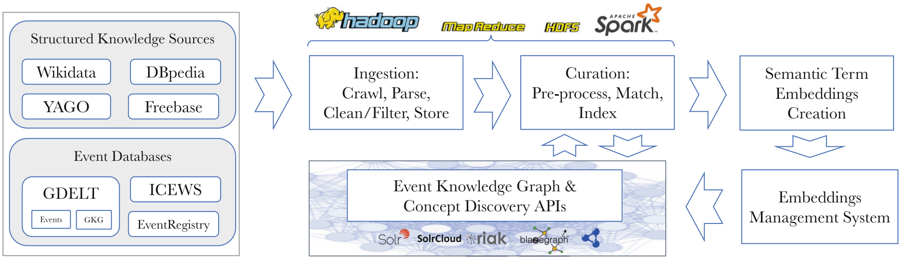


Fig. 3. System architecture

GDELT. The Global Data on Events, Location, and Tone (GDELT) project [17] claims to be "the largest, most comprehensive, and highest resolution open database of human society ever created". GDELT data contains three databases. GDELT Event database provides *coded* event data based on a popular scheme using the CAMEO (Conflict And Mediation Event Observations) coding framework [24] to code actors and actions. GDELT includes other features of the events such as the date of the event, information about the source articles, numerical scores reflecting the "tone" of the source articles and other similar features, and geographical coordinates. The second database provided by GDELT is the Global Knowledge Graph (GKG) and contains records describing the source articles of the events. Each record provides a comprehensive set of numerical features for the article, in addition to annotations with several dictionaries of persons, organizations, and "themes". The third database GDELT provides is the Mentions database which connects event records with GKG article records. The most recent version of GDELT data is updated daily and includes historical data since February 2015. At the time of this writing, we have ingested 128,946,346 Event records, 157,082,264 GKG records, and 436,961,965 Mention records.

ICEWS. Integrated Conflict Early Warning System [11,30] provides a coded event database similar to the GDELT Event database. ICEWS event records describe features of source and target actors including their name, "sector", and country, features of the action including date, source, a short text description, and geographical descriptions. A recent version of the data also includes CAMEO codes for actions. We have ingested the most recent publicly available data that has a coverage of historical events from 1995 to 2015, with 14,757,915 records.

EventRegistry. The EventRegistry [16] project takes a completely different approach than the coded event databases, and performs event extraction based on a clustering of news articles and event mentions. EventRegistry records contain a multilingual title and summary text, the number of articles reporting the event, the event date (when the event has happened or will happen and not the report date as in coded event databases), and a set of concepts along with the concept type (e.g. location, person, or "topic") and its Wikipedia URL. At the time of this writing, we have ingested 2,889,497 event records extracted from 33,549,078 news articles from the past two years, with 98,435,900 concept annotations, 42,006,079 similarity links, 772,553 location annotations.

Knowledge Sources. In addition to event data, our system also ingests publicly available RDF knowledge bases to use as a source of reference knowledge. Our current knowledge sources include Wikidata [29], DBpedia [8], YAGO [23], and Freebase [9]. At the time of this writing, we have ingested over 6.3 billion RDF triples, containing over 488 million entities (unique URIs) and over 83 million English label statements.

2.2 Ingestion

As shown in Fig. 3, we have a common ingestion pipeline for both the event databases and knowledge sources that is capable of crawling remote sources, parsing structured relational, semi-strcutured (JSON), and RDF (NTriples) data, cleaning invalid records or statements and applying basic filters (e.g., removing non-English labels), and finally storing the data. Our platform is implemented on top of Apache Hadoop and Spark, enabling efficient data processing on a cluster on public or private cloud.

2.3 Curation

We adopt a pay-as-you-go integration approach [13,19] and perform only a minimal curation by a lightweight mapping of known entities, linking them using a common URI when possible. To integrate the knowledge sources we use the existing Wikipedia URLs. We then index all the facts (RDF triples) in a key-value store (powered by Riak [4]) in addition to a document store (powered by SolrCloud [5]) that makes it possible to perform highly efficient fact-based or label-based lookups. We also create an auxiliary unified index of common entities using our mapping strategy that results in a collection of 16,108,676 entities with Wikipedia URLs, each linked with one or more of their Wikidata, DBpedia, YAGO, and Freebase URIs. All the event databases are indexed in a similar way in our key-value and document stores, with labels matched and linked with a Wikipedia URL when possible.

2.4 Semantic Embeddings Engine

Inspired by the idea of word embeddings in NLP [21], recent work has proposed the use of shallow neural networks to map values in structured tables to vectors (referred to as embeddings) [10]. This enables powerful semantic similarity queries even without a prior knowledge of the database schema and contents. We adopt a similar strategy and transform every value in the input event databases into an embeddings vector using a variation of the *continuous skip-gram model* of the original *word2vec* [7,20,21]. The first step in this process is a *virtual document* creation process in MapReduce, turning each row in the input database into a context in a corpus of text. We then feed the text corpus into a word2vec model construction modified to take into account the different characteristics of structured data:

- The order of columns in structured databases is of little importance. While distance between two words in a text document makes them farther in terms of context, the first column in a database table is as relevant to the second column as to the last column.
- In text documents, typically a random-sized window of words is selected. The length of each database record is fixed and so there is no need for a random window size.
- Most importantly, while all words in a text corpus are treated in the same way and do not have specific roles, values in different columns in structured sources describe different (event) features and may need to be grouped and queried differently. There is often a need to search over (or query using) the terms from specific attributes (columns).

Once attribute values are mapped into low-dimensional vectors, aggregate vectors can represent individual records (articles or events), and similarity queries over the vectors can be used for concept discovery and analysis as described in Sect. 3. These vectors represent the semantic context of every single value seen in the input data, enabling a powerful and extremely efficient method of performing similarity analysis over large amounts of data. As an example, the corpus size (number of words in the "virtual documents") for GDELT GKG is 23,901,358,498 while the size of the vocabulary (number of unique words) in our embeddings is 2,829,213. Still, a key requirement is efficient similarity queries over the vectors with milliseconds running time to enable real-time analysis queries through our UI (Figs. 1 and 2) as some analysis queries require several similarity queries each over millions of vectors. We achieve this using the efficient Annoy library [1] as the core of our embeddings management system.

2.5 Event Knowledge Graph and Concept Discovery APIs

The final outcome of all the components is a set of APIs to perform knowledge graph and concept discovery queries. In particular:

- **Lookup APIs.** These APIs provide access to the ingested and curated event data and knowledge. For example, one can perform search over knowledge base entity labels and subsequently retrieve human-readable facts as JSON objects. Using this API the user can retrieve infobox-style information about each of the concepts shown under the "Global Context" box in Fig. 2a. These APIs also enable queries across event databases, e.g., retrieve ICEWS, GDELT, and EventRegistry events in a given time range that is annotated with a particular concept.

- **Natural Language and Keyword Query Understanding APIs.** These APIs turn the user query into a set of knowledge base concepts and event database terms. In Fig. 2a, the concepts shown under the "Global Context" are extracted using the API that outputs knowledge base concepts, whereas the terms shown under "DeepSim Context" are terms found in GDELT GKG data used for the shown concept discovery results.
- **Concept Discovery and Ranking APIs.** These APIs take a set of concepts or terms and return as output a ranked list of concepts of different types (e.g., Persons, Organizations, Themes). Details of the concept ranking algorithms are described in the following section.

3 Concept Ranking Algorithms

In this section, we describe three classes of algorithms for concept discovery and ranking. These algorithms *identify* and *rank* a set of most relevant concepts of various types (e.g., persons or organizations) for a given set of concepts. An example use of these rankings is shown in Fig. 2 where sorted lists of ranked concepts relevant to the user's analysis question are shown. The end goal is providing the output either directly to an analyst or to other components of an analysis system. We first describe the algorithms, followed by an evaluation of their effectiveness in identifying relevant concepts.

3.1 Index-Based Method (`co-occur`)

The `co-occur` method relies on an efficient index to measure the level of co-occurrence of concepts in a collection of events and uses this as a measure of relevance. Using the index described in Sect. 2.3, we can search for (all or recent) event records annotated with a given set of concepts. By counting the concept annotations for every record in the output, a list of most frequently co-occurring concepts of various types is returned along with the percentage of co-occurrence of the annotations among all the retrieved event records. Figure 2b shows an example of ranked "Topic", "Key Player" (Person), and "Location" concepts over EventRegistry event records. The concepts extracted from the input question are "Caracas", "Protest", and "Venezuela". Obviously, these concepts themselves are on top of the lists as they appear in 100% of the event records containing them. The topic concept "Government" appears in 87% of the events and "Nicolás Maduro" appears in 69% of the events, indicating that these concepts are highly relevant to the input concepts in recent events.

3.2 Deep Similarity Method Using Semantic Embeddings (`context`)

The `context` method relies on the term embeddings built over an event database as described in Sect. 2.4. First, a vector is retrieved for each of the terms extracted from the input question (where there exists a vector representation in the embeddings space), and an average vector is constructed by summing the values in each dimension and normalizing the resulting vector. Using the embeddings management system, the most similar vectors of various kinds of terms are retrieved, ranked by their similarity to the average vector. Figure 2a shows an example of concept rankings with the same question. The API used in the following evaluation queries embeddings built over 157 million GDELT GKG records, with vectors of size 200 and cosine similarity as our choice of vector similarity measure. These rankings result in less-obvious and harder-to-explain but deeply relevant sets of concepts in the output.

3.3 Combination Methods

We implement two combination methods. In the `co-occur_context` method, we retrieve a set of 3*k results of an index-based method, re-rank the output using the embeddings-based similarity of the terms in the output, then select the top k terms. In the `context_co-occur` method, we retrieve 3*k results of the `context` retrieval and sort the output based on their position in the `co-occur` results before selecting the top k terms. In the following section in Table 1, we show an example of how these re-rankings improve the results.

4 Experiments

To evaluate the performance of the concept ranking algorithms, we use queries represented as sets of query terms that would be extracted from an analyst's question. We sought to identify a 'ground truth' of concepts related to each query, where a concept is a person or organization directly involved with the topic of the query.

As mentioned in Sect. 2, our framework enables real-time or near real-time response time for each of the algorithms and so we do not compare running times.

4.1 Evaluation Data

To our knowledge, there is no existing public data set that identifies key people and organizations for a set of analytical questions. To provide an objective basis for our evaluation, we used reports that summarize a political or social event or situation, making the assumption that these reports are a response to such a question, and will mention the most important key players. We did not use news articles because these are the source of GDELT events, so as not to bias the results. We identified three potential sources of reports:

- Declassified US Government intelligence reports. We were only able to identify one such report that relates to the time period covered by the GDELT GKG event database: 'Assessing Russian Activities and Intentions in Recent US Elections', released in January 2017.

- Wikipedia pages describing a newsworthy event or topic with relevance to social unrest. For example 'Impeachment of Dilma Rousseff' or 'Shortages in Venezuela'.
- Human Rights Watch reports. These are detailed descriptions of specific human rights situations around the world, for example 'Philippine Police Killings in Duterte's "War on Drugs"'. 1,091 such reports are available in HTML.

Using these sources we developed test queries consisting of a small set of query terms and 'ground truth' sets of people and organizations. The query terms can include a country, people, organizations, and themes (drawn from the GDELT GKG themes described earlier). The 'ground truth' items are selected from the people and organizations mentioned in the report, to represent the ideal response of the system to the query. Three query sets were developed: *Manual*, *Curated*, and *Auto*.

Manual. A small set of 12 hand-built queries derived from 6 Wikipedia pages, 5 Human Rights Watch Reports from 2014 or later; and 1 declassified intelligence report. All queries specified the country most strongly associated with the report, and 1–3 manually selected themes from GDELT GKG, in addition to any people or organizations mentioned in the report title. For example, the query terms for the Wikipedia page 'Impeachment of Dilma Rousseff' consisted of the country 'Brazil', the person 'Dilma Rousseff', and the theme 'IMPEACHMENT'.

The ground truth concepts were the people and organizations mentioned in each report as key players for the topic. We removed concepts that were found in the report but not in our embedding. Subsequently, each query had an average of 10 ground truth people and 7 ground truth organizations.

Curated. A set of 25 queries based on Wikipedia pages describing events from 2014–2016, where the query terms (country, people, organizations and themes) were selected manually, but the ground truth terms were automatically generated from the Wikipedia links within the page, and then curated to remove non-person and non-organization terms. Some people and organizations mentioned in the original report may be missed in this process.

Auto. A larger set of 179 queries derived from the Human Rights Watch reports, with fully automatic generation of both query terms and ground truth. To generate the query terms, the query builder used the document title, subtitle and teaser - a short paragraph of a few sentences describing the report. It used concept extraction software that combines output from ClearNLP [3] and OpenNLP [2] to identify noun phrases referring to named entities, and assigns types to them according to their linguistic context. We relied on these types to identify countries, people and organizations in the text. The ground truth people and organizations were generated by using the same concept extraction software, applied to the full text of the report. We removed people and organizations not found in our embedding. Finally, we selected the 179 queries that had a country, at least one other query term (person, organization or theme), and had ground truth terms that were not already present in the query. Of these, the majority (102) were queries consisting of a country and the single organization 'Human Rights Watch'. 26 contained a person, and 51 contained an organization other than Human Rights Watch. From these, we further selected only the usable queries with ground truth items that were not in the query (143 for people and 155 for organizations) These queries had, on average, 21 ground truth people and 32 ground truth organizations.

4.2 Example Results

Table 1 shows the set of ground truth people and the output of the algorithms for a query from the manual test set, based on the 2016 Human Rights Watch report "Venezuela's Humanitarian Crisis. Severe Medical and Food Shortages, Inadequate and Repressive Government Response"[1].

Of the 14 most relevant people identified in the report, only 7 were present in the embedding (indicated in column 1 with (*)). Two of the others were not found in the GKG data at all, and the remaining five were mentioned only 1–59 times - not enough to be included in the embedding. For organizations, 17 were mentioned in the report, and 9 of these were found in the embedding. The GKG data does not often include common acronyms like BBC or FBI, although there are some exceptions. This creates challenges for automated testing since the reports often use an acronym to refer to an organization.

The most relevant people mentioned in the report, 14 in all, are listed in the first column of Table 1, while the remaining columns show the top 14 results for each algorithm, given the query for the country "Venezuela" and the GKG theme "SELF_IDENTIFIED_HUMANITARIAN_CRISIS". The seven items from the ground truth that are potentially findable in the index and in the embedding are indicated with (*) in column one, while the found items are highlighted in bold in the subsequent columns, including alternate spellings of the same person's name.

The `co-occur` method finds only the Venezuelan leaders in the top 14. It also returns ten other world leaders, politicians and spokespeople. These people have either made statements about Venezuela's humanitarian crisis, or Venezuela has made comments about their own country's crisis (e.g. Bashar Assad). Although Donald Trump was not yet president of the United States during the period covered by the data, his opinions on foreign policy in Latin America were discussed in the news, and he made statements about the situation in Venezuela. Two journalists who write frequently about Venezuela are also suggested (Joshua Goodman, Gonzalo Solano).

In marked contrast, the `context` method's results do not include any foreign leaders and politicians. Instead, there are seven Venezuelan politicians, along

[1] https://www.hrw.org/report/2016/10/24/venezuelas-humanitarian-crisis/severe-medical-and-food-shortages-inadequate-and.

Table 1. Example results from each algorithm for the query "Venezuela", "SELF_IDENTIFIED_HUMANITARIAN_CRISIS". (*) indicates those candidates that were potentially findable in the GKG data.

Ground truth	`co-occur`	`context`	`co-occur_context`
Nicolás Maduro (*)	**Nicolás Maduro**	**Delcy Rodriguez**	**Delcy Rodriguez**
Hugo Chávez (*)	Barack Obama	**Nicholás Maduro**	**Nicholás Maduro**
Zeid Ra'ad Al Hussein (*)	Rafael Correa	Jesus Torrealba	Hannah Dreier
Ban Ki-moon	**Hugo Chávez**	Vladimir Padrino	**Luis Almagro**
Delcy Rodríguez (*)	Joshua Goodman	Henrique Capriles	Juan Manuel Santos
Luisana Melo (*)	John Kerry	**Nicolás Maduro**	Barack Obama
Luis Almagro (*)	Bashar Assad	Jorge Arreaza	Rafael Correa
Johan Gabriel Pinto Graterol	Donald Trump	Hannah Dreier	**Hugo Chávez**
Julio León Heredia	Juan Manuel Santos	Girish Gupta	Joshua Goodman
Carlos Zapa	Gonzalo Solano	Eyanir Chinea	John Kerry
Flor Sánchez	Vladimir Putin	Andrew Cawthorne	Bashar Assad
Diosdado Cabello (*)	David Granger	David Smilde	Donald Trump
Rafael Uzcátegui	John Kirby	Ernesto Villegas	Gonzalo Solano
Feliciano Reyna	Salva Kiir	**Luis Almagro**	Vladimir Putin

with four journalists and a human rights advocate and academic (David Smilde), and the secretary-general of the Organization of Latin American States (Luis Almagro). Some of these politicians are very closely associated with the humanitarian crisis in Venezuela, notably Vladimir Padrino, the Venezuelan Minister of Defense, who is responsible for food distribution, even though they were not mentioned by name in the report.

Combining these methods by ranking the first 90 `co-occur` results according to their `context` ranking moved five highly related candidates to the top of the list, including a new ground truth person: Luis Almagro. Similarly, the `context_co-occur` method (omitted from Table 1 for space constraints) moved four items to the top of the ranking, including the misspelling of Nicolás Maduro as Nicholás Maduro. Both combination methods slightly increased the number of ground truth items found in the top 10 ranked results from 2 or 3 to 4 out of a possible maximum of 7.

4.3 Evaluation Method

To evaluate and compare the methods of identifying key players, we applied each of the four methods (`co-occur`, `context`, `co-occur_context` and `context_co-occur`) to the test query data sets (manual, curated and auto), for both people and organizations. All methods were limited to 30 returned candidates. For each query, we calculated four classic information retrieval evaluation measures: precision (ratio of correct concepts in the output), recall (ratio of ground truth concepts returned in the output), F1 (harmonic mean of precision and recall), and average precision (average precision value at all the ranks where a correct concept is returned). Overall values for each test set are reported as the mean of the values for the individual queries in the set. Following the recommendation by Smucker et al. [25], we performed randomization test and two-tailed paired samples t-tests to test for statistical significance.

4.4 Evaluation Results

Manual. Table 2 shows the results for the manual data set. For person experiments, all measures showed better performance from the combination methods, with the context method performing the lowest. The co-occur_context method outperformed the co-occur method by 19%. However, pairwise comparisons of F1 scores between methods showed only the (context,context_co-occur) and (co-occur,co-occur_context) pairs to be statistically significantly different ($p < 0.05$). For the organization experiments, again the co-occur_context combination method performed best over all four measures, but only the (context,context_co-occur) pair was found to be statistically significant in terms of comparison by MAP or F1 scores. The lack of statistical significance is due to the high variance of the results for each query, and show in part the need for a larger data set for a proper comparison as our overall results described in Sect. 4.4 also confirm.

Table 2. Accuracy results over the manual data set.

	Person				Organization			
	co-occur	context	co-occur context	context co-occur	co-occur	context	co-occur context	context co-occur
MAP	0.233	0.199	0.251	0.233	0.179	0.143	0.184	0.189
F1	0.192	0.174	0.228	0.213	0.178	0.107	0.183	0.141
Pr.	0.133	0.121	0.158	0.149	0.117	0.066	0.119	0.089
Re.	0.372	0.328	0.437	0.388	0.436	0.304	0.459	0.374

Curated. Table 3 shows the results for the curated data set. For the person experiments, the overall pattern was very similar to the manual data set, with the co-occur_context method showing the best performance across all measures, including an 18% improvement for F1 over the co-occur method. For F1, the differences between the (context,co-occur), (context,context_co-occur) and (co-occur,co-occur_context) pairs were statistically significant. For the organization experiments the co-occur and co-occur_context methods performed the best, and their F1 scores were not significantly different, while all other pairwise comparisons were, except for the two lowest performing methods: context and context_co-occur.

Table 3. Accuracy results over the curated data set.

	Person				Organization			
	co-occur	context	co-occur context	context co-occur	co-occur	context	co-occur context	context co-occur
MAP	0.132	0.070	0.135	0.123	0.130	0.039	0.075	0.051
F1	0.140	0.104	0.165	0.143	0.142	0.058	0.142	0.058
Pr.	0.119	0.090	0.139	0.124	0.107	0.042	0.108	0.045
Re.	0.251	0.160	0.300	0.225	0.290	0.116	0.289	0.122

Auto. Table 4 shows the results for the auto data set. For these results, organizations followed a similar pattern to the two other datasets, and all pairwise comparisons were statistically significant for all metrics, with the only exception for MAP, where the two combination methods were not distinguishable. For person experiments, the results were lower, less than 0.1 for all metrics and methods, so that while the combination methods produced around 10% higher average scores, the differences were not statistically significant, with the exception of the (`context`,`context_co-occur`) and (`co-occur`,`co-occur_context`) pairs for F1 or MAP.

Table 4. Accuracy results over the auto data set.

	Person				Organization			
	co-occur	context	co-occur context	context co-occur	co-occur	context	co-occur context	context co-occur
MAP	0.041	0.046	0.050	0.051	0.132	0.073	0.117	0.116
F1	0.058	0.060	0.066	0.066	0.165	0.084	0.157	0.108
Pr.	0.051	0.056	0.059	0.061	0.173	0.088	0.163	0.112
Re.	0.090	0.086	0.099	0.094	0.224	0.114	0.217	0.150

Comparing Results Across the Data Sets. We also explored whether the different data sets provided similar results. Figure 4 shows F1 values for people (left) and organizations (right) as boxplots. Each box indicates the interquartile range of the data, the center line indicates the median value, the whiskers above and below give the 95% confidence intervals, and circles indicate outliers. Significant differences are indicated above with red brackets. For people, the less curated sets of queries produced lower results, but the pattern of results is very similar across all three datasets. Recall that the auto queries did not contain any themes, and so they often did not capture the topic of a report well, giving the system a low chance of success. Results were twice as good for organizations as for people in the auto data set, probably reflecting the large number of queries that included an organization. Again, the pattern of results remained similar across the data sets. We observed similar trends for other accuracy measures.

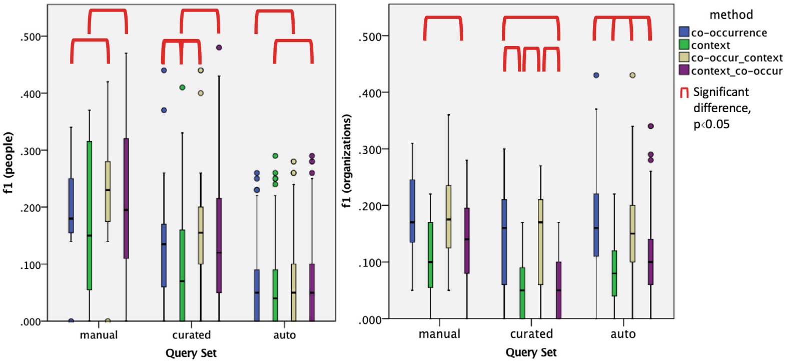


Fig. 4. Results for people and organizations across the data sets and four methods

4.5 Discussion

Overall, both the combination algorithms performed better than the individual `co-occur` and `context` algorithms. This suggests that combining methods did go some way towards addressing the individual weaknesses of the two approaches, with an effect size of up to 19% improvement over the best individual method.

Not surprisingly, the algorithms produced the best results on the manual test set, followed by the curated set, and the lowest values for the automatically generated set, which has less well constructed queries that do not capture the topic of the report as well. Importantly, the similarities between data set results when comparing the concept discovery algorithms increases confidence in the evaluation, and more generally in the use of automated methods as a valid and scalable way to approach the evaluation of concept discovery algorithms, despite the noise and loss of accuracy compared to hand-curated data.

Our approach to evaluation has some limitations. Our source reports do not mention all of the people and organizations relevant to the topic by name. We do not translate mentions like “The Minister for the Interior” into a named person, and nor do we attempt to resolve references to groups like “Brazilian steel companies.” Our methods also draw from articles published after the publication of the report, when new concepts may be introduced. All people or organizations found by our methods but not named in the ground truth are treated as wrong answers, but some of these may be highly relevant to the topic. In the example shown in Table 1, the majority of the persons returned by the `context` method are in fact highly relevant despite the fact that our input report did not contain their names. This shows that a potential use case for our system is complementing analysts in finding concepts that are not already covered in their report. Also note that there is often a major difference between the number of candidates proposed (30) and the number of ground truth items provided (generally less than 30). Thus, our reported accuracy scores are very low and underestimate the overall quality of the responses.

Another way to evaluate our framework would be to compare the output of the algorithms with that of human analysts. We did not take this approach because our goal is to surface potentially unexpected concepts, helping analysts to mitigate cognitive bias. Thus we preferred an objective evaluation method.

5 Conclusion and Future Work

In this paper, we presented a framework for discovering concepts related to an analysis question using event databases. We showed how the use of semantic technologies enable a unified query mechanism across event databases using mappings to general-domain knowledge bases, and semantic embeddings. We then presented three classes of concept ranking algorithms and performed an evaluation of the quality of the rankings using a corpus of reports written by humans. We plan to extend our ground truth data sets for concept discovery, in part using crowd sourcing and human analysts, and make the outcome publicly available for future research. Our plan is to use our benchmark as a way to compare different event databases such as ICEWS, GDELT, and EventRegistry, and highlight their strengths and shortcomings. Finally, we are planning to analyze how the outcome of concept discovery affects other analysis tasks such as scenario planning [26] and building forecast models [15].

References

1. Annoy. https://github.com/spotify/annoy. Accessed 8 May 2017
2. Apache OpenNLP (v1.5.3). https://opennlp.apache.org/. Accessed 18 May 2017
3. ClearNLP (v3.2.0). https://github.com/clir/clearnlp. Accessed 18 May 2017
4. Riak. https://github.com/basho/riak. Accessed 8 May 2017
5. SolrCloud. https://wiki.apache.org/solr/SolrCloud. Accessed 8 May 2017
6. TAC KBP 2016 Event Track. https://tac.nist.gov/2016/KBP/Event/index.html. Accessed 8 May 2017
7. Word2vec: tool for computing continuous distributed representations of words. https://code.google.com/archive/p/word2vec/. Accessed 8 May 2017
8. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - a crystallization point for the web of data. JWS **7**(3), 154–165 (2009)
9. Bollacker, K.D., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: SIGMOD, pp. 1247–1250 (2008)
10. Bordawekar, R., Shmueli, O.: Enabling Cognitive Intelligence Queries in Relational Databases using Low-dimensional Word Embeddings. CoRR abs/1603.07185 (2016). http://arxiv.org/abs/1603.07185
11. Boschee, E., Lautenschlager, J., O'Brien, S., Shellman, S., Starz, J., Ward, M.: ICEWS Coded Event Data (2017). https://doi.org/10.7910/DVN/28075
12. Doddington, G., et al.: The automatic content extraction (ACE) program tasks, data, and evaluation. In: LREC, May 2004
13. Franklin, M.J., Halevy, A.Y., Maier, D.: From databases to dataspaces: a new abstraction for information management. SIGMOD Rec. **34**(4), 27–33 (2005)

14. Hogenboom, F., Frasincar, F., Kaymak, U., de Jong, F., Caron, E.: A survey of event extraction methods from text for decision support systems. Decis. Support Syst. **85**(C), 12–22 (2016). https://doi.org/10.1016/j.dss.2016.02.006
15. Korkmaz, G., Cadena, J., Kuhlman, C.J., Marathe, A., Vullikanti, A., Ramakrishnan, N.: Combining heterogeneous data sources for civil unrest forecasting. In: ASONAM, pp. 258–265 (2015). https://doi.org/10.1145/2808797.2808847
16. Leban, G., Fortuna, B., Brank, J., Grobelnik, M.: Event registry: learning about world events from news. In: WWW, pp. 107–110 (2014)
17. Leetaru, K., Schrodt, P.A.: GDELT: global data on events, location, and tone, 1979–2012. In: ISA Annual Convention (2013)
18. Lin, D., Pantel, P.: Concept discovery from text. In: COLING, pp. 1–7 (2002)
19. Madhavan, J., Jeffery, S.R., Cohen, S., Dong, X., Ko, D., Yu, C., Halevy, A.: Web-scale data integration: you can only afford to pay as you go. In: CIDR (2007)
20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)
22. Muthiah, S., et al.: Embers at 4 years: Experiences operating an open source indicators forecasting system. In: KDD, pp. 205–214 (2016)
23. Rebele, T., Suchanek, F.M., Hoffart, J., Biega, J., Kuzey, E., Weikum, G.: YAGO: a multilingual knowledge base from Wikipedia, Wordnet, and Geonames. In: ISWC, pp. 177–185 (2016)
24. Schrodt, P.A., Yilmaz, O., Gerner, D.J., Hermreck, D.: The CAMEO (conflict and mediation event observations) actor coding framework. In: 2008 Annual Meeting of the International Studies Association (2008)
25. Smucker, M.D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: CIKM, pp. 623–632 (2007)
26. Sohrabi, S., Riabov, A., Katz, M., Udrea, O.: An AI planning solution to scenario generation for enterprise risk management. In: AAAI (2018)
27. Sohrabi, S., Udrea, O., Riabov, A.V., Hassanzadeh, O.: Interactive planning-based hypothesis generation with LTS++. In: IJCAI, pp. 4268–4269 (2016)
28. Tan, A.H.: Text mining: the state of the art and the challenges. In. In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases, pp. 65–70 (1999)
29. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
30. Ward, M.D., Beger, A., Cutler, J., Dickenson, M., Dorff, C., Radford, B.: Comparing GDELT and ICEWS event data. Analysis **21**, 267–297 (2013)