Translating Narratives into Knowledge Graphs

Andrea Westerinen and William Westerinen
OntoInsights, LLC
64 Woodholme Way, Elkton, MD 21921
USA

andrea@ontoinsights.com, jeff@ontoinsights.com

ABSTRACT

Currently available analysis tools for text processing are often solely based on statistical, machine learning (ML) methods. Many of these tools, while powerful for detecting patterns, lack the ability to fully access and analyze the underlying meanings embedded in narrative text. To address this shortcoming, an approach (termed Deep Narrative Analysis, DNA) has been prototyped. DNA supplements ML by pairing it with semantic, linguistic, and ontological offerings and insights. It creates graphical, semantically-rich interpretations of free-form text, allowing more complexity and nuance to be captured and to inform analyses.

Keywords*: Narratives, Ontology Development, Natural Language Parsing Natural Language Understanding*

1.0 INTRODUCTION

This paper overviews the Deep Narrative Analysis (DNA) architecture and prototype and its potential use in infrastructures examining textual sources such as autobiographical narratives, blog posts and news articles. DNA builds on machine learning-based textual parsing to generate knowledge graphs, and then uses the graphs to perform detailed comparisons and analyses. For example, one of DNA's objectives is to enable text-analysis applications to highlight where (and with what word choices) sources describe people and situations of importance, such as election candidates, as well as political or healthcare debates. This level of information can, in turn, inform researchers regarding the emotions and memories that are intended to be triggered in the readers.

DNA is currently an early phase research project with potential application in multiple fields. The underlying technologies (parsing narratives, creating graphical representations of the texts, fusion of external reference material, and narrative comparison and analysis) are demonstrable today. More advanced analysis of the narratives using machine learning and pattern analysis are being implemented.

In this paper, the focus is on several main topics: how natural language processing is performed and augmented by linguistic and semantic insights to produce knowledge graphs from text (Section 3.2 and 3.3); and examples of various machine learning and pattern recognition techniques to discover specific concepts and repeated patterns in the narratives (Section 3.4). An example analysis is discussed in Section 4. The paper begins by discussing the rationale behind the work (Section 2) and ends with a discussion of next steps, potential areas of application of the technology and conclusions (Sections 5 through 7).

2.0 RATIONALE

Technologies based on Artificial intelligence (AI) and Machine Learning (ML) have had significant success. Large Language Models (LLM) such as OpenAI's GPT-3[1] and Google's LaMDA[2] are in the headlines, able to support human-like conversations. These offerings are, indeed, useful for narrative analysis. However, on their own, they lack even the rudimentary ability to process or understand the underlying meanings embedded within most texts. By relying on what are fundamentally statistical methodologies (along with huge corpuses of data), these LLMs are limited in their ability to process and/or act on semantic factors (such as underlying or implied meanings).

2.1 DNA and Semantic Analysis

It is our position that ML-based text processing is necessary but not sufficient to provide deep understanding of text. Indeed, it is excellent for:

- Part of speech tagging
- Finding relationships between words/clauses of a sentence (as found with dependency parsing)
- Named entity recognition of persons, organizations, places, events, dates, etc.
- Correlation of words (such as accomplished using GPT-2 and -3)

DNA's purpose is to augment LLM and other ML-based applications with semantic, linguistic, and rules-based reasoning to enable a deeper analysis of text-based entities (as well as those that can be converted to text).

There are three general areas where DNA-based applications will have functionality that is generally lacking in LLMs (and other ML-based applications) [11]:

- Cognitive modelling
- Referencing (the outside world)
- Compositionality

2.1.1 Cognitive Modelling

Humans make sense of the real world using cognitive models. For example, two individuals will likely have the same cognitive model of a car. It allows them to communicate about cars in general or about a specific instance. In addition, the individuals can infer and draw conclusions based on the model – e.g., if a car that was working yesterday is not starting today, they might reason that it is out of fuel. When applying ML techniques to text-based sources, this type of conclusion is difficult to reach without such a model.

In a similar fashion, DNA has its own cognitive model in the form of a backing ontology which is used to create and structure the graphical representations of narratives. Using the ontology to encode events discussed in the narratives enables comparison and reasoning. Hypothesis testing can be performed (e.g., the belief that the car is

[1] https://openai.com/api/

[2] https://blog.google/technology/ai/lamda/

out of gas), contradictions can be investigated (such as finding that the car was actually not 'working yesterday' due to a traffic accident), etc. This is described in more detail in section 3.4, below.

2.1.2 Referencing

The second factor necessary for semantic understanding of text is relating the words to the (real) world. For example, if the word "Paris," is in a narrative, it is probably referencing a city in the country of France, at a certain latitude and longitude. LLMs would recognize "Paris" as being statistically associated with the word "France" and perhaps "Louvre" and probably "Seine" but would have no notion about them beyond a statistical correlation.

DNA, on the other hand, can connect words within a narrative to the world and supplement the information with online details (for example, by accessing and linking to external sources such as Wikidata[3] and Geonames[4]). This not only provides additional meaning but enables DNA-based applications to display additional information to users (plotting locations on a map, providing a description for a referenced business or organization, and much more).

2.1.3 Compositionality

LLMs can create realistic sentences that are readable but lack real coherence across the sentences. For example, in the real world, the left rear wheel of a car is connected to the left axle that is, in turn, connected to the differential (and so forth). However, when asking an LLM to construct a sentence about a car's wheels, it may (inadvertently) imply that the wheel is connected directly to the differential. The LLM has no knowledge of how a car is constructed – only of the words that are statistically related to cars and their wheels.

DNA's ontology provides applications with the means to describe people, places, times, events and conditions, and their relationships, including compositionality.

[3] https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service

[4] https://www.geonames.org/export/geonames-search.html

3.0 DESCRIPTION: DEEP NARRATIVE ANALYSIS ARCHITECTURE

This section introduces the DNA architecture, which is shown in Figure 1.

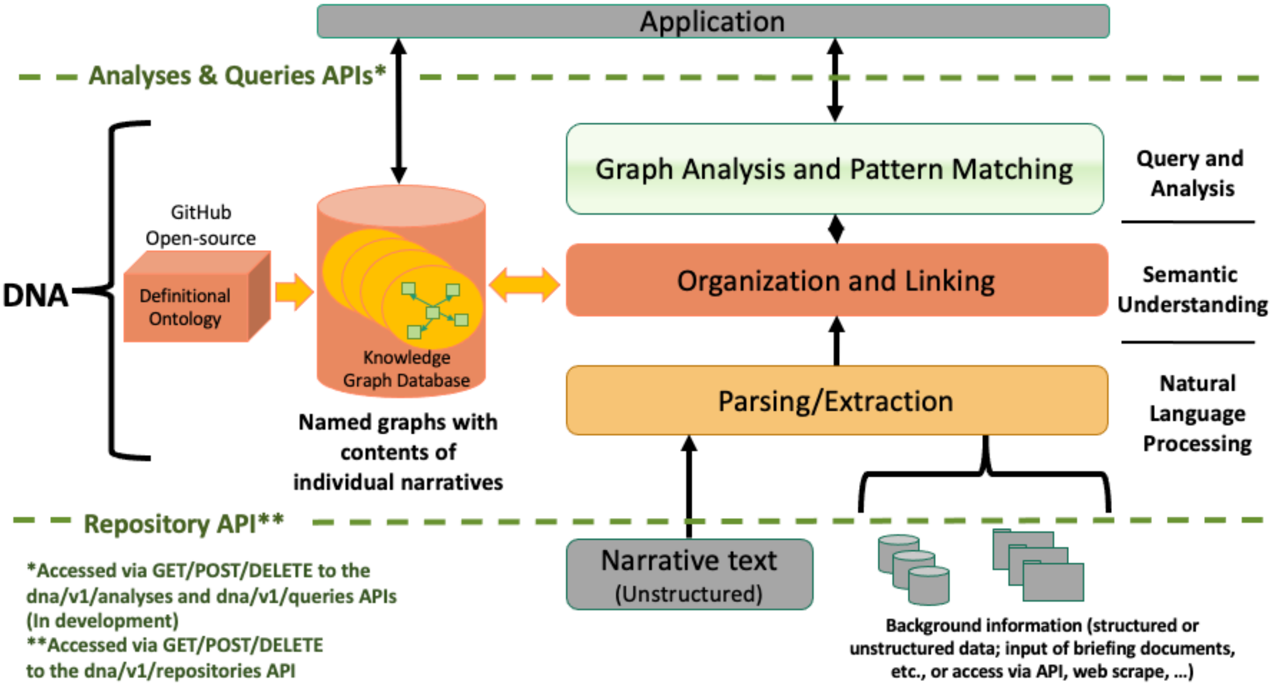

Application
Analyses & Queries APIs*
DNA
GitHub Open-source
Definitional Ontology
Knowledge Graph Database
Named graphs with contents of individual narratives
Graph Analysis and Pattern Matching
Organization and Linking
Parsing/Extraction
Query and Analysis
Semantic Understanding
Natural Language Processing
Repository API**
*Accessed via GET/POST/DELETE to the dna/v1/analyses and dna/v1/queries APIs (In development)
**Accessed via GET/POST/DELETE to the dna/v1/repositories API
Narrative text (Unstructured)
Background information (structured or unstructured data; input of briefing documents, etc., or access via API, web scrape, ...)

Figure 1. DNA High Level Architecture

3.1 Overview

Work on DNA has primarily been focused on the “Parsing/Extraction” and “Organization and Linking” steps of the architecture. Access to these components is provided via a set of RESTful APIs (the dna/v1/repositories API [17]) supporting the manipulation of data stores and the ingest, update and removal of narratives from a store. The underlying technologies are discussed in more detail in the sub-sections below.

Using DNA, each sentence of a narrative is extracted, parsed, and transformed into a graphical representation, based on Semantic Web standards[5] and structured with the semantics of the DNA ontology (discussed in more detail in Section 3.3). Similarly, background material (such as a glossary of actors, locations, times, and events) may also be ingested. The background information could be scraped from a web page, extracted by DNA from a briefing document, manually input, accessed via an API, or similar means.

[5] https://en.wikipedia.org/wiki/Semantic_Web

After ingest and transformation, the various narrative graphs can be compared/contrasted, examined for similarities and differences, analyzed based on network structure and different measures, and more.

Although the “Graph Analysis and Pattern Matching” functionality has not been fully implemented, our first stage analyses and visualizations are demonstrated using a Jupyter notebook[6] (which is included in our GitHub repository). In the future, the functionality will be accessible via the dna/v1/analyses and dna/v1/queries APIs.

The last layer of the architecture, the application layer, is outside the scope of the current work.

3.2 Use of Linguistic Event Theory

DNA’s ontology and graphical representation of source material are focused on the “verbs” in the texts - capturing what is described, what has happened, and what is experienced. This focus, in turn, leads to a different approach to ontology development.

Oftentimes, when an ontology is developed, the work emphasizes the “things” (e.g., nouns) in the domain of interest. Indeed, in Stanford's paper, "Ontology Development 101: A Guide to Creating Your First Ontology" [16], the authors say, "Classes are the focus of most ontologies. Classes describe concepts in the domain." In the Stanford paper, the classes are nouns - wines and food.

For DNA, however, our approach is centered on verbs (encoding events and conditions) as the core concepts of the ontology. This approach is based on research into storytelling (using the events relayed in a story to explain who we are and how our lives have evolved [22]) and consistent with linguistic event theory [23]. The latter views an event as an individual entity (a “thing”) that can have relationships to other entities such as the agents participating in the event, instruments used, locations and times where and when the event occurred, etc. Events can be decomposed into lower-level sub-events, have properties, and be connected in temporal, spatial and cause/effect relationships.

To understand the difference in focus, a typical (non-DNA based) extraction from text (as exemplified by [2] and [3]) encodes a sentence’s subjects and objects as entities that are related. For example, consider the sentences, “George lived in Detroit in 1980. He moved to Atlanta in 1990.” In this example, Person and Location (with individuals, George, Detroit, and Atlanta) are the main concepts, and they are connected via a relationship, “lives in.” A problem arises, however, related to describing when each relationship holds (e.g., living in Detroit versus Atlanta). The timing would appear to be a property of the relationship, “lives in.”

Using the event-based approach of DNA, the number of individuals and the relationships change, and they become more standardized. Now, there are four main concepts – Person, Location, the condition indicated by the verb phrase “lives in” (DNA’s Residence class) and the event of moving. When encoding the example sentences, there is an occurrence of Residence (George living in Detroit) and an instance of the Movement event, which then can be used to infer a new instance of Residence. All the events have a relationship to the same agent (George). One has a location relationship referencing Detroit and a time relationship indicating 1980. The other two have a location of Atlanta and the times, 1990 and “after” 1990. There is no ambiguity

[6] https://jupyter.org/

regarding how time is captured (it is a property of the events), and comparisons and analyses across instances of Residence are straightforward. Patterns of residence can be analyzed. Alignment across the events happening in a Location or to a Person can be performed.

3.3 The DNA Ontology

As discussed in the previous section, DNA's ontology is focused on events, situations and the people and things experiencing or involved in them. The goal is to use the ontology to describe the "6Ws" (who, what, where, when, why and how). The top-level, core ontology is shown in Figure 2 and illustrates how the ontology relates to the "6Ws."

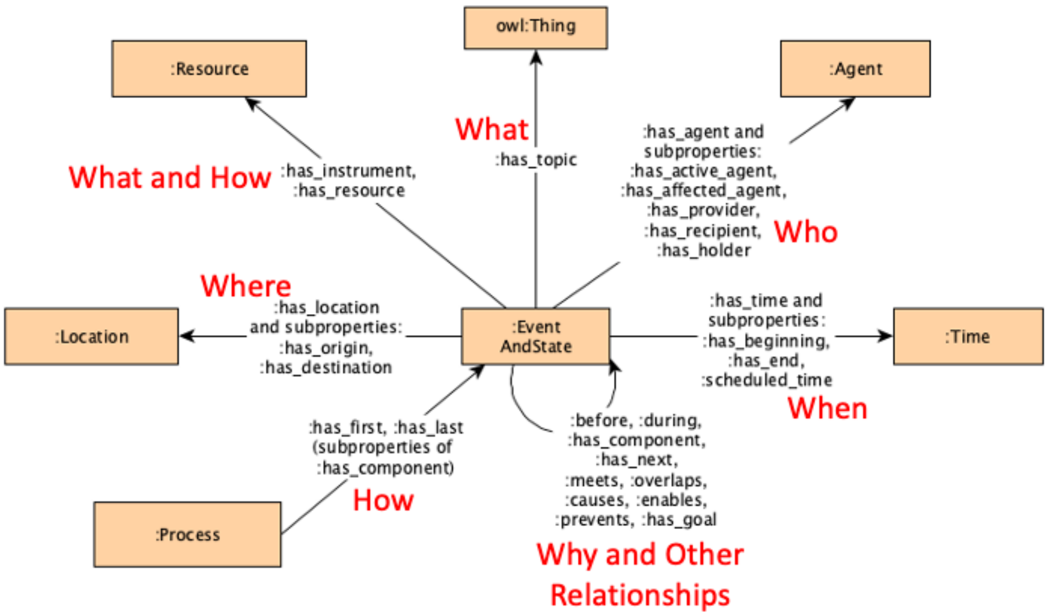


Figure 2. DNA Top-Level Ontological Concepts

Existing taxonomies, lexical resources and ontologies were used in the development of the DNA ontology. Specifically:

- The Resource class hierarchy is based on the UN Harmonized System commodity codes[7]
- The Time classes reuse concepts from the W3C Time ontology[8]
- The Location class hierarchy is inspired by the GeoNames Feature Classes and Codes[9]

[7] https://www.foreign-trade.com/reference/hscode.htm

[8] https://www.w3.org/TR/owl-time/

[9] https://www.geonames.org/export/codes.html

- The EventAndState classes are based on the Conflict and Mediation Event Observations (CAMEO) codes [21], Automatic Content Extraction (ACE) Guidelines for Events [9], event nugget and sequencing concepts from the Text Analysis Conference's Knowledge Base Population (TAC KBP) workshops [14] and other Event ontologies (such as [1], [7] and [20]), as well as incorporating a list of the most common English verbs[10]

In addition, all of the ontology concepts have been supplemented using the synonym details of WordNet [18]. This was accomplished by first manually mapping the DNA concepts into WordNet's noun and verb synonym sets, and then programmatically traversing the WordNet hyper-/hyponym trees for additional terms.

3.4 Natural Language Processing

DNA's analyses rely on the creation of a "knowledge graph" to encode the occurrences described in text. A knowledge graph can be defined as:

> *A collection of entities (the "nodes") representing specific instances of the types of things in a domain of interest, interconnected by named "edges" which identify the entities' properties (such as string or integer values) and relationships between them. The structure and semantics of the graph ("knowledge") are derived from the use of an ontology.*

To create the graph, DNA follows a logical sequence based on first chunking the sentences to obtain and individually evaluate subject-verb-object subclauses. For example, the sentence, "When Mary went to the grocery store, John practiced guitar," would be separated into two parts - "Mary went to the grocery store" and "John practiced guitar." This is done to improve the resolution of the sentence parse.

After chunking, the clauses are analyzed using the part of speech and dependency parse components of the spaCy open-source, NLP library [6]. The root verb of each clause is extracted, along with the subject, object, and prepositional details. All the semantically related words that make up those noun, verb and prepositional phrases are captured. For example, using the subclauses from the Mary and John sentence above, the verbs, went and practiced, are extracted as well as the nouns, Mary, grocery store and guitar.

Once each subject-verb-object (or preposition) phrase is parsed, rules and heuristics are applied that utilize lexical and grammatical knowledge to determine the "most likely" semantics. Continuing the Mary and John example, Figure 3 shows the resulting event details that would be encoded in a knowledge graph.

[10] https://www.wordexample.com/list/most-common-verbs-english/

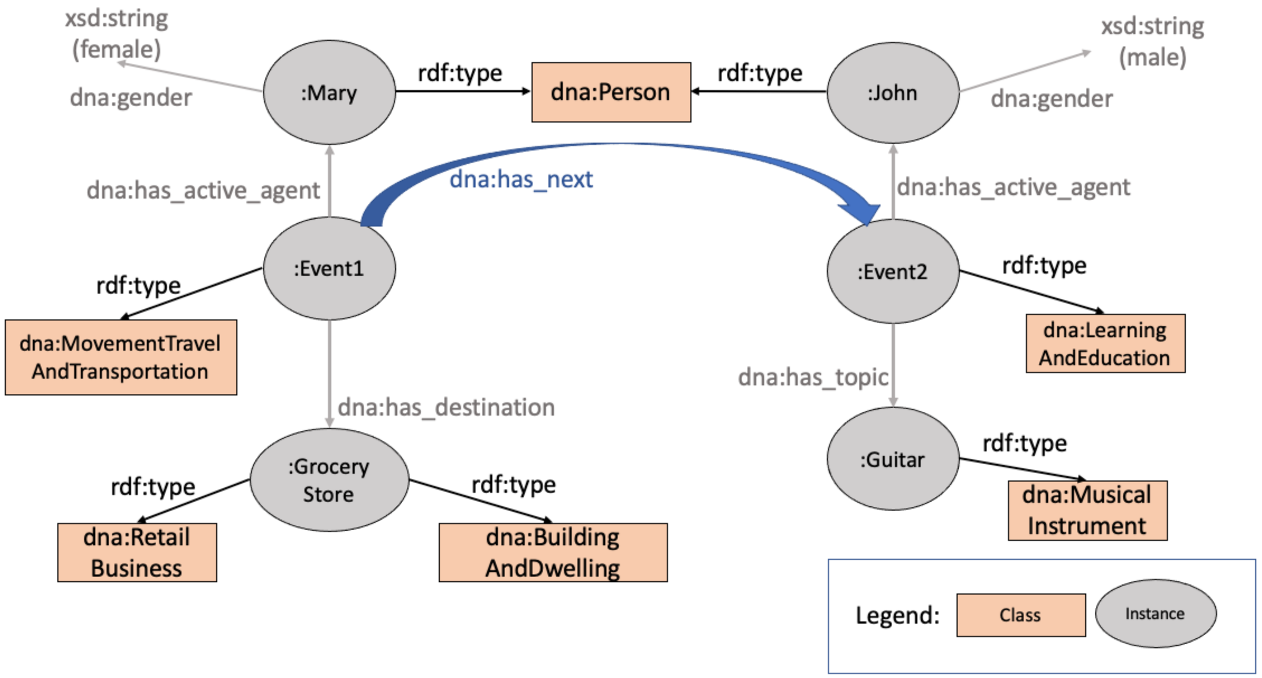


Figure 3. Event Details Encoded Using the DNA Ontology

3.5 Graph Analysis

Support for complex queries and the addition of new knowledge (via inference-based reasoning) are two motivators for using ontology-backed knowledge graphs.

Queries can be written to extract information such as the persons or locations discussed in texts, the types of events that are described and with what words, the persons or organizations that frequently publish and/or utilize key phrases or memes, the types and number of quotations used in the texts, and much more. In addition, the queries can support grouping by different criteria (for example, right- or left-leaning biases) to compare and contrast results.

Although the current DNA prototype directly queries the data store using the SPARQL[11] semantic web language (a query language similar to SQL), a user-friendly front-end is being designed that utilizes templated queries and a question-answering interface. The front-end accesses the data store using the dna/v1/queries API.

Support for reasoning to infer new knowledge and to validate the consistency of the text parses are other motivating factors for using knowledge graphs. As one example of "new knowledge", consider the childhood rhyme, "Jack and Jill went up the hill, to fetch a pail of water." This implies that there is a well or other water

[11] https://www.w3.org/TR/sparql11-query/

source on top of the hill. And, with a broader set of background/contextual information (topographical descriptions of the area showing the hill, and real estate data indicating that a house is built there which is connected to the area's water system), you could conclude that the water specifically came from a spigot and not a well.

Using open-source libraries such as NetworkX[12], the characteristics and structure of the extracted knowledge graphs can be compared. The possible analyses include graph density and centrality values, location and comparison of sub-graph patterns, discovery of cliques (groups of people or organizations who reference each other), analysis of how nodes (events) are connected in a text and much more. This enables visualization of the various focuses of a narrative, news article or blog post – addressing whether the text emphasizes a single topic or discusses a variety of different and (potentially) unrelated information.

One scenario that is possible when analyzing a broad set of articles is to use the centrality concept[13] to detect probable instances of "super-spreading". Based on centrality measures, the graphs derived from many different sources could indicate which authors, references or quotations recur. Once identified, they could be investigated further to examine and potentially validate their role as catalysts for "super-spreading" of messages.

Given that the DNA ontology is used to reduce text to a series of semantically-rich events, entity alignment is possible. This enables the comparison of texts, based on the events that are mentioned, the flow of the sentences, the people and sources that are referenced, and the specific word choices that are used.

Examples of many of these analyses and comparisons are shown for a set of 4 news articles in Section 4.

3.6 Related Work

This section overviews existing work in the use of ontologies for news analysis and natural language understanding. One of the key motivators for the creation of ontologies is to describe and enable reasoning about domains of interest. An important "domain of interest" is the analysis of news articles. Using an ontology for this purpose is not new. References to a variety of design approaches were provided in Section 3.3. A paper by Liu, et. al. [10] provides a valuable survey of current research.

Going further, however, use of an ontology to aid in the specific task of propaganda analysis is a relatively new area of work. It was proposed by Hamilton [5], where she discusses the motivations for

> *"the development of an ontology for the media ecosystem with the explicit purpose of detecting propaganda techniques and rhetorical devices in the news... This ontology can then be used to build a knowledge base from news articles".*

[12] https://networkx.org/

[13] https://en.wikipedia.org/wiki/Centrality

Although the overall characteristics and use cases for the “media ecosystem” ontology are presented in Hamilton’s paper, no concrete ontology is defined. DNA provides an example of a potential backing ontology and a prototype infrastructure that uses it.

Research taking a similar, hybrid (ML-, AI- and semantics-based) approach to textual analysis includes work by Gary Marcus (mentioned in Section 2) and Marjorie McShane ([12]). McShane’s and Nirenburg’s book, “Linguistics for the Age of AI” [13], discusses the concept of a “language-endowed intelligent agent” (LEIA) which is able to “understand, explain and learn” (DNA is focused on the first of these capabilities – understanding).

According to McShane and Nirenburg, text must be transformed to “an ontologically grounded text meaning representation.” They describe 7 steps to accomplish natural language understanding:

- Pre-semantic analysis addressing pre-processing and semantic parsing of the text
- Pre-semantic integration organizing and providing deeper syntactic insights via dependency parsing and the use of various heuristics
- Basic semantic analysis mapping the extracted text to the ontology
- Basic coreference resolution resolving textual ambiguities as much as possible
- Extended semantic analysis adding information from external sources and via reasoning “to improve the semantic/pragmatic analysis of not only individual sentences but also multisentence discourses”
- Situational reasoning based on situational awareness, common-sense knowledge, plans, goals, etc.

DNA provides a preliminary infrastructure addressing the first six steps and is being extended to address the last bullet item. For a more detailed understanding of how the DNA prototype implements these concepts, we recommend consulting DNA’s open-source code base[14], where one can find details related to each of the steps.

4.0 EXAMPLE EXTRACTION AND COMPARISON

In order to better understand the functionality of DNA, an exemplary analysis is performed on a small set of news articles. These articles were chosen as representative of center-, right-and left-leaning publications based on University of Michigan Library’s Research Guide on “fake news,” lies and propaganda [24].

Four articles were chosen for the analysis:

- Centrist (Wall Street Journal, WSJ): Liz Cheney Concedes to Trump-Backed Challenger in Wyoming Primary[15]

[14] https://github.com/ontoinsights/deep_narrative_analysis

[15] https://www.wsj.com/articles/liz-cheney-faces-uphill-fight-in-primary-against-trump-backed-opponent-11660642201 (Note that only the first 11 paragraphs were used for quotation and name analysis since the remainder of the article discussed Cheney’s political career)

- Far-right leaning (Breitbart): Liz Cheney Compares Herself to Abraham Lincoln in Concession Speech[16]
- Right leaning (Fox News): Rep. Liz Cheney compares herself to Abraham Lincoln following resounding defeat in Wyoming primary[17]
- Left leaning (New York Times, NYT): Liz Cheney Invokes Lincoln and Grant in Impassioned Concession Speech[18]

Specific details regarding DNA's approach to natural language understanding were discussed in Section 3. The focus of this section is to demonstrate how the output of the syntactic and semantic processing (knowledge graphs) can be used in analyses. Note that this is a small use case to show the detail provided by the DNA approach, at a scale that can be reviewed by the reader.

For the purposes of demonstration, only the first five paragraphs of the news articles were parsed and analyzed. However, in two cases, the articles were fully analyzed to extract:

- All mentions with people's names
- All quotations if the quotation had at least a subject and verb

4.1 Narrative Query

Performing SPARQL queries across the knowledge graphs that encode texts' semantics is the basis for story arc visualization and alignment (discussed later, in Sections 4.3 and 4.4), which are key features of DNA. The results of queries are also generally useful as a straightforward means to gather data for visualizations (such as histograms) and for text processing tasks (versus writing and debugging custom code).

The following queries demonstrate several simple text comparisons that are significantly more straightforward based on the DNA encodings of the parses.

DNA uses the Stardog Free Knowledge Graph[19] as its backing data repository, and programmatically interfaces with it using the pystardog Python library[20]. Queries are written using the SPARQL query language as was discussed in Section 3.5.

An example of a SPARQL query is shown in Figure 3, with its results. The query is searching for quotations that were used in multiple articles. It examines two quotations (located by checking for quotations that are associated with a 'narrative'/news article) and returns them as the variables, ?text1 and ?text2. It also returns the publisher name associated with those narratives as the variables, ?publisher1 and ?publisher2. The query includes a "filter" to make sure that an instance is not compared with itself (line 7 of the query), and filters if

[16] https://www.breitbart.com/politics/2022/08/16/liz-cheney-compares-herself-abraham-lincoln-concession-speech/

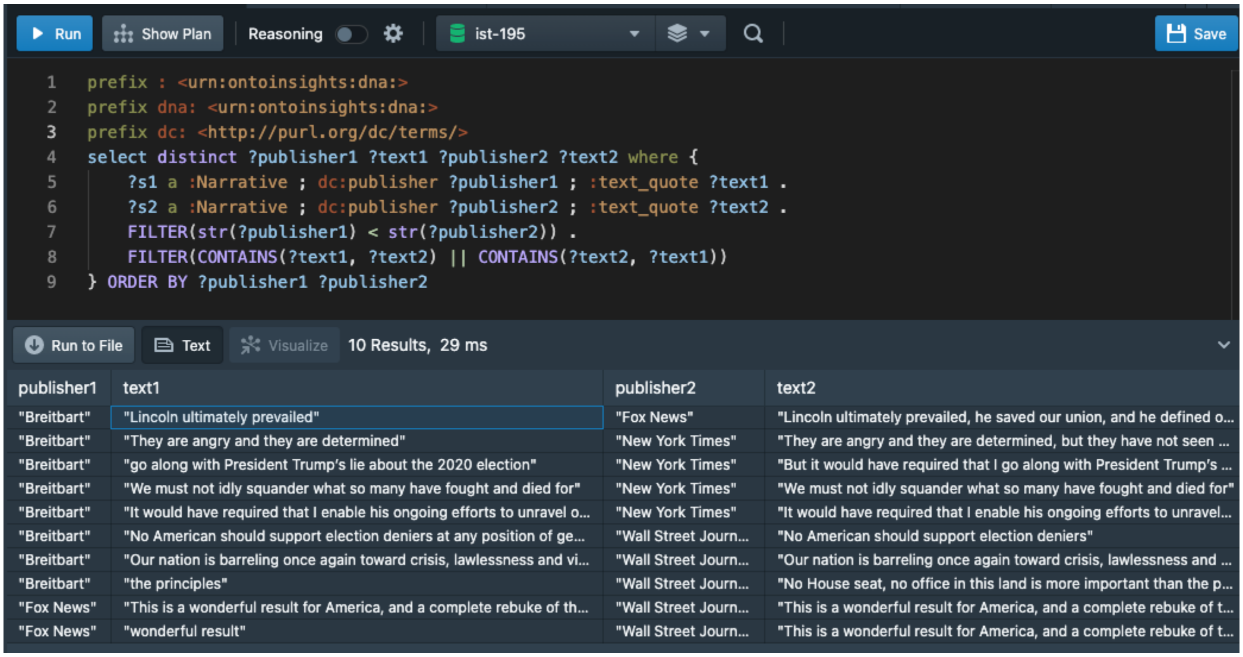
[17] https://www.foxnews.com/politics/liz-cheney-compares-herself-abraham-lincoln-resounding-defeat-wyoming-primary

[18] https://www.nytimes.com/2022/08/17/us/politics/liz-cheney-concession-speech.html

[19] https://www.stardog.com/platform/

[20] https://pystardog.readthedocs.io/en/latest/source/stardog.html#module-stardog.connection

any quotation text is contained in another (line 8). It is important to check the texts in both directions (text1 contained in text2, or text2 contained in text1) since writers do not necessarily quote full sentences.



```
1  prefix : <urn:ontoinsights:dna:>
2  prefix dna: <urn:ontoinsights:dna:>
3  prefix dc: <http://purl.org/dc/terms/>
4  select distinct ?publisher1 ?text1 ?publisher2 ?text2 where {
5      ?s1 a :Narrative ; dc:publisher ?publisher1 ; :text_quote ?text1 .
6      ?s2 a :Narrative ; dc:publisher ?publisher2 ; :text_quote ?text2 .
7      FILTER(str(?publisher1) < str(?publisher2)) .
8      FILTER(CONTAINS(?text1, ?text2) || CONTAINS(?text2, ?text1))
9  } ORDER BY ?publisher1 ?publisher2
```

10 Results, 29 ms

publisher1	text1	publisher2	text2
"Breitbart"	"Lincoln ultimately prevailed"	"Fox News"	"Lincoln ultimately prevailed, he saved our union, and he defined o...
"Breitbart"	"They are angry and they are determined"	"New York Times"	"They are angry and they are determined, but they have not seen ...
"Breitbart"	"go along with President Trump's lie about the 2020 election"	"New York Times"	"But it would have required that I go along with President Trump's ...
"Breitbart"	"We must not idly squander what so many have fought and died for"	"New York Times"	"We must not idly squander what so many have fought and died for"
"Breitbart"	"It would have required that I enable his ongoing efforts to unravel o...	"New York Times"	"It would have required that I enable his ongoing efforts to unravel...
"Breitbart"	"No American should support election deniers at any position of ge...	"Wall Street Journ...	"No American should support election deniers"
"Breitbart"	"Our nation is barreling once again toward crisis, lawlessness and vi...	"Wall Street Journ...	"Our nation is barreling once again toward crisis, lawlessness and ...
"Breitbart"	"the principles"	"Wall Street Journ...	"No House seat, no office in this land is more important than the p...
"Fox News"	"This is a wonderful result for America, and a complete rebuke of th...	"Wall Street Journ...	"This is a wonderful result for America, and a complete rebuke of t...
"Fox News"	"wonderful result"	"Wall Street Journ...	"This is a wonderful result for America, and a complete rebuke of t...

Figure 3. Query Example

As can be seen in the results, the Breitbart article uses the same quotes as in the other articles. However, there is little overlap of the choice of quotations across the other texts. Performing further analysis on the types of quotations that are repeated by far-right and far-left publications could provide valuable insights into the use of social media to influence readers.

Queries can also be used to directly generate statistics for visualization. As noted above, once the text is encoded in the form of a knowledge graph, querying the graph and visually displaying the results (with the help of plotting libraries) becomes extremely straightforward. Figures 4 and 5 were generated from query results using the matplotlib[21] library. Figure 4 visualizes how often Liz Cheney, Dick Cheney, Donald Trump and others are specifically mentioned in the article texts. Figure 5 shows how many quotations occurred in each of the articles.

At the level of four articles, the results are interesting. Again, the Breitbart article appears to "name drop" far more frequently than the others. But, scaling this type of analysis across a large number of articles is needed.

[21] https://matplotlib.org/

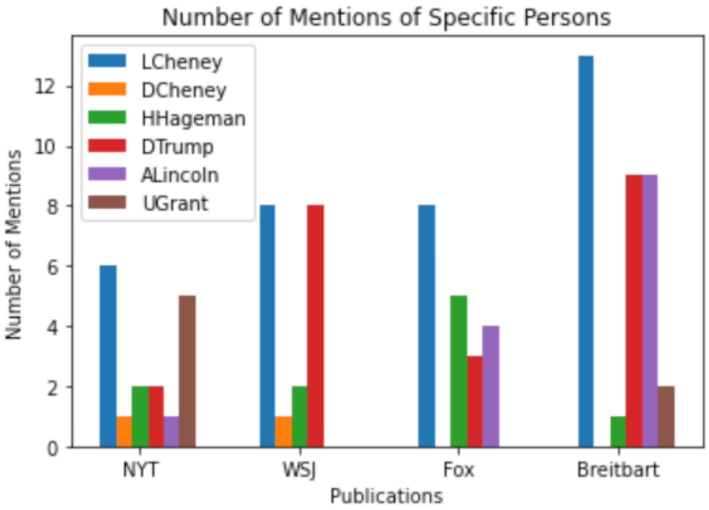


Figure 4. Number of Named References to Individuals

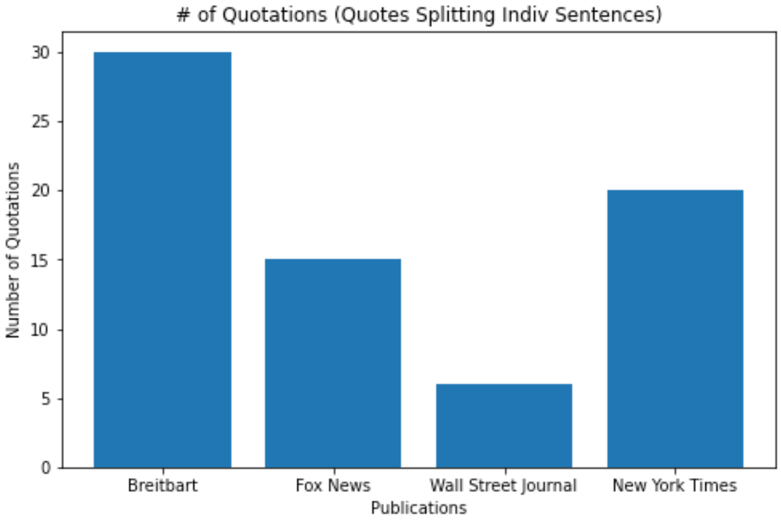


Figure 5. Total Number of Quoted Sentences in the Articles

4.2 Knowledge Graph Characteristics, Measures and Similarity

As mentioned in section 3.5, the concept of centrality is important in network analysis. Centrality is an indication of “key” nodes in a network (typically ones having the most connections passing through them). It would indicate which entities (such as events or actors) are the most referenced in an article or blog post. Alternately, across a large set of articles, centrality, and other graph measures such as community (clique) detection could be used to discover which themes, persons, memes, etc. are repeated (spreading through the network). Having a straightforward means to calculate these is valuable.

Centrality (as well as other graph measures) is easily generated using the DNA knowledge graphs. The DNA encodings of the articles were converted to a NetworkX format (discussed in Section 3.5), and then various measurements were reported. Table 1 shows the results of the analyses over the first five paragraphs of each article (with the removal of the quotation details except when present in those paragraphs).

Source	# Nodes	# Edges	# Nodes with a Single Edge	Degree Centrality Mean, Std Dev	Node with Highest Degree Centrality
NYT	93	101	60	0.0236, 0.0232	:Liz_Cheney
WSJ	102	111	66	0.02155, 0.0211	Event noting Harriet Hageman win
Fox	113	131	74	0.0207, 0.02188	:Liz_Cheney
Breitbart	82	98	52	0.0295, 0.0338	Event noting that Cheney failed to “go along with” Trump

Table 1. Comparison of Graph Measurements

The results above are reasonable for the NYT, WSJ and Fox, since the articles are about Liz Cheney’s concession speech (and Harriet Hageman’s win). However, the Breitbart article’s highest centrality node is related to the sentence, “Cheney claimed she lost her primary election only because she failed to ‘go along with President Trump’s lie about the 2020 election.’”

Analysis of a larger set of articles (and the full text of the articles) would be needed to draw higher confidence conclusions or insights. This is ongoing work at the present time.

4.3 Story Arc Visualization

Analysis of a large corpus of news articles, forum texts and social media posts can yield additional insight into trends, inchoate super-spreading events and memes, as well as early identification of new and upcoming influencers. However, these sources may contain complex sentences which can be difficult to parse manually, or may require with custom code or text pre-processing. DNA is designed to extract the relevant details from text (working syntactically outward from the root verb) and encode and summarize those details to create a story arc for each article. The arc is valuable to highlight the main details and flow of a narrative. Using it could aid in better understanding a narrative.

Figure 6 shows the story arc extracted from the Wall Street Journal article. The text was ingested and processed as described in Section 3.4, and the semantics of each sentence were captured in one or more event declarations. (Multiple events could be generated from a sentence due to chunking.) When the event is encoded, both the full text and a summary "label" for each chunk are associated with it. The chunk labels are displayed in the visualization.

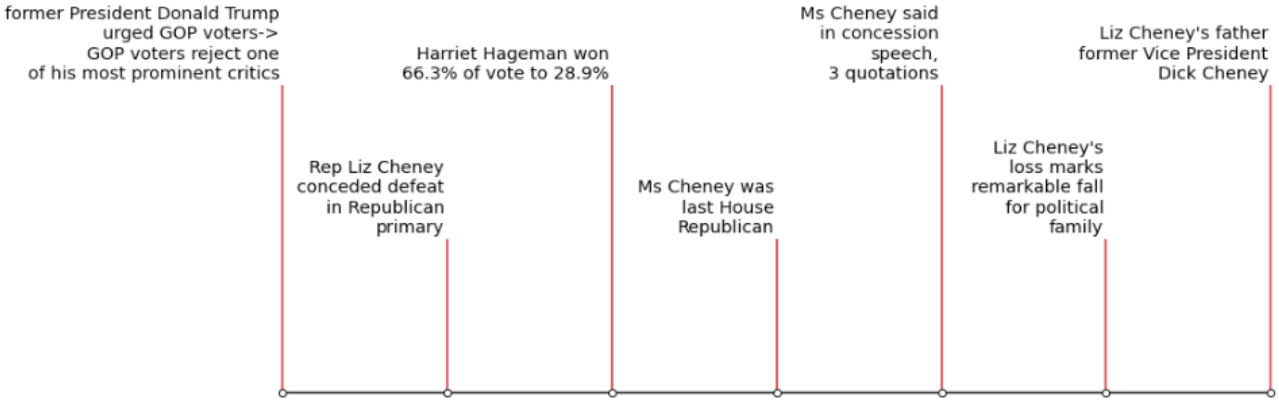


Figure 6. Story Arc from the WSJ Article

Examining the figure, a few things should be noted:

- When the label includes an arrow (->), one can assume that the root verb of the sentence had a modifying clause which included both a subject and verb (so it was chunked as a new sentence)
  - An example of this occurs in the first label from the first sentence of the article. The text indicates that Liz Cheney's defeat was a Trump priority and that he "urged GOP voters" to reject Cheney.
  - From this sentence, the first label captures the "urging" and the intended "rejection" events. The second label captures Cheney's "defeat".
- The fourth item on the story arc is an example where human-in-the-loop intervention would be valuable
  - The item should be deleted or modified by a user
  - The top-level parse and encoding are relatively correct (Liz Cheney has an affiliation to the Republican party) but the overall parse could not make sense of the details
  - The original sentence in the text was "Ms. Cheney was the last House Republican to face a primary among the 10 who voted to impeach Mr. Trump for his actions related to the Jan. 6 Capitol riot."
- When the text includes a quotation with its own subject and verb, the quotation is retained but removed from the parse
  - Item 5 indicates that three of Cheney's sentences were quoted (and are easily retrieved) but are not included in the parse
  - This will be a configurable option in a future release of the dna/v1/repositories API

The full text of the WSJ article is repeated here, for comparison.

> *U.S. Rep. Liz Cheney conceded defeat Tuesday in the Republican primary in Wyoming, an outcome that was a priority for former President Donald Trump as he urged GOP voters to reject one of his most prominent critics on Capitol Hill.*
>
> *The race wasn't close. Harriet Hageman, a water and natural-resources attorney who was endorsed by the former president, won 66.3% of the vote to Ms. Cheney's 28.9%, with 95% of all votes counted.*
>
> *Ms. Cheney was the last House Republican to face a primary among the 10 who voted to impeach Mr. Trump for his actions related to the Jan. 6 Capitol riot.*
>
> *"No House seat, no office in this land is more important than the principles we swore to protect," Ms. Cheney said in her concession speech. "Our nation is barreling once again toward crisis, lawlessness and violence. No American should support election deniers."*
>
> *Ms. Cheney's loss also marks a remarkable fall for a political family that has loomed large in Republican politics in the sparsely populated state for more than four decades. Her father is former Vice President Dick Cheney, who was elected to the House in 1978, where he served for a decade.*

One final item should be noted. The first sentence of the second paragraph is not included in the story arc. It was ignored by DNA since the sentence's subject was not a person, an object was not specified, and the verb did not connote any action.

4.4 Story Alignment

Entity alignment can be a valuable technique for comparing how various sources provide coverage of one or more events. For example, alignment analysis across news articles, forums and social media posts is critical to understanding where and how different narratives overlap and where they diverge. This is not possible with classic machine learning techniques, and it would be extraordinarily difficult for custom code to provide this functionality without a means to obtain the backing semantics. DNA can offer a semantic understanding of what is discussed in text-based articles, which can be the basis of aligning events that are reported differently in various sources.

Table 2 shows the results of a query across the knowledge graphs of the four news articles searching for event correspondences. Events were assumed to correspond/align if they were instances of the same event class, and referenced the same person as the subject or object.

Demonstrating this functionality at scale (for a large number of articles) is currently in development.

Event	NYT	WSJ	Fox	Breitbart
Cheney loss/defeat in primary	Sentence 8	Sentence 1, Sentence 8	Sentence 1, 6	Sentences 1, 2
Cheney comparison to Lincoln	(After initial 5 paragraphs)	(Not present)	Sentence 1	Sentences 1, 6
Percentage Spread	Sentence 8	Sentence 3	(Not present)	(Not present)
Loss to Hageman; Hageman win	Sentence 8	Sentence 3	Sentence 1, 6	(Not present)

Table 2. Story Alignment

Regarding where the texts diverge, the discussion of Liz Cheney's win in the 2020 primary (in the NYT article) and Trump's "pure delight" over Cheney's loss (in the Breitbart article) are two examples.

Development of a visualization strategy that clearly articulates the alignment is in-progress.

4.5 Other Features

DNA can provide the following additional visualizations and analyses of text-based sources:

- Event detection and enumeration – Since DNA builds on Linguistic Event Theory (section 3.2), a list of events can be extracted from text sources, enumerated, and presented for additional processing and analysis
- Actor(s) detection and enumeration – A consolidated list of influencers, authors, and other persons can be extracted and presented
- Timeline analysis – In articles that have definite timelines (such as a description of historical events over a given timeframe), a consolidated timeline of events, actors and locations can be presented to users
- Location analysis – Given that DNA has access to location information and can extract geographical information from text (using named entity recognition), a graphical mashup based on a mapping application (such as Google Maps) can be presented to a user, showing how events and actors are distributed geographically

Demonstration and/or proofs of concept for each of these items constitutes ongoing work items for DNA.

5.0 FUTURE WORK

DNA's current architecture (including its rules-based processing of narrative sentences) is published in full, as open-source, as was mentioned in Section 3.6. However, there is much more research and implementation to be done, including:

- Support for non-English text
    - DNA's code base has been architected to extend beyond English through its use of spaCy (which currently supports over 20 languages) and WordNet[22]; Several, alternative languages are currently being investigated
- Enhanced grammatical analysis to deal with constructions such as clauses within clauses and support for complex conjunctive structures
- Extended analysis by clustering texts based on their author or authoring organization, use of specific phrases and memes, reference to the same (or similar) authoritative sources and/or reuse of the same/similar quotations, etc.
    - With identification of the earliest reference and a timeline of propagation
- Enhanced graphical analyses, as discussed in Section 3.5
- Improved entity alignment, making use of more powerful algorithms and toolkits (as described in [4] and [25])
    - With better scalability; For example, handling the ingest of a large number of stories simultaneously, as well as the ability to run more powerful queries
- Incorporation and use of common-sense knowledge, which is typically assumed and not explicitly stated in text
    - A recent survey paper [8] overviewed the most popular sources of common-sense knowledge, and classified them across 13 dimensions (such as whether they address synonymy, antonymy, meronomy, etc., whether they classify topics via a subsumption hierarchy, and whether they describe the utility of an object or the desires and goals of an agent)
    - Aligning the concepts based on the DNA ontology, from several semantic common-sense knowledge bases, will improve the insights derivable from the texts
        - For example, common-sense knowledge can be used to predict conditions, motivations and emotions before, during or after events discussed in a text, or it can be used to predict the overall emotion that should be generated in the reader
        - Two relevant common-sense knowledge bases are the Atlas of Machine Commonsense for If-Then Reasoning (ATOMIC [19]) which enhances reasoning about events by describing their pre- and post-states, and Generalized and Contextualized Story Explanations (GLUCOSE [15]) which captures causal theories regarding the progression of events in a narrative
- Deployment of a cloud-based implementation (e.g., hosted on AWS)

[22] http://globalwordnet.org/resources/wordnets-in-the-world/

6.0 FUTURE APPLICATIONS

DNA has the potential to support a wide variety of applications across many domains. The dis-mis-information domain is addressed in Sections 6.1 and 6.2. A list of other potential areas where DNA technology could be applied is briefly overviewed in Section 6.3.

6.1 Narrative Analysis from News Sources, Forums and Social Media Feeds

An application area that will benefit from the inclusion of DNA technology is for dis-/mis-information analysis. DNA is designed to be a component of a toolkit analyzing text from on-line news sources and forums. This section discusses that use case in more detail.

Typically, a user or researcher is interested in specific topic areas and/or publishing organizations, forums or people. Or, they might be interested in discovering what is being written by select actors. DNA's line-by-line sentence analysis could be applied to texts meeting their criteria/interests. As noted in Section 3.5, the characteristics and structure of the extracted knowledge graphs could be directly compared. Cliques (groups of people or organizations who reference each other) could be discovered. Various visualizations of the texts' knowledge graphs could be presented. And, with a broader set of articles, the graphs could be used to indicate which authors or quotations act as catalysts for "super-spreading" of messages. Also, DNA could be used to provide measurements (such as graph density or centrality) that aid in the detection of propaganda and "fake news."

A possible workflow for this application would begin with a source selection screen, where news articles, social media posts, forums and other on-line (and off-line) material can be searched for, selected, and retrieved or uploaded. Once the user selects the source material, the relevant texts would be ingested and processed. Output could take the form of consolidated timelines, lists of events, actors and locations detected in the text, or details related to the patterns or cliques discovered.

Figure 7 shows a possible process flow for the application. As can be seen from the diagram, we believe that given the state of the art of NLU and NLP, there is a need for human-in-the-loop to add corrections and remove ambiguities that occur in the parse and ontology mapping (and, of course, there will need to have an interface developed that facilitates this). As the toolset and backing linguistic and semantic data matures, human-in-the-loop will become less critical. However, for high precision results (e.g., in legal cases, versus for introductory analyses), there will likely always be a need for an interface to edit the knowledge graphs that are used in the analyses (even once the prototype evolves into a production toolset).

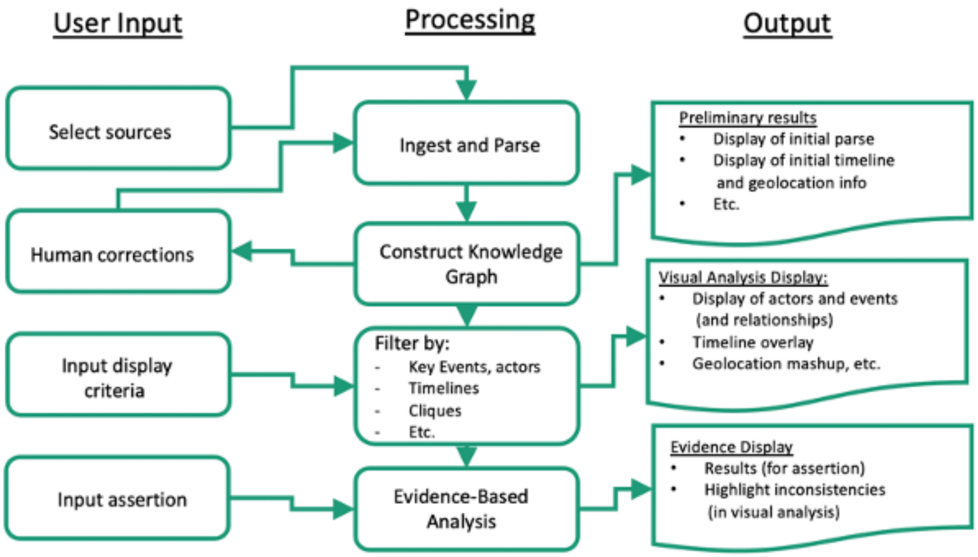
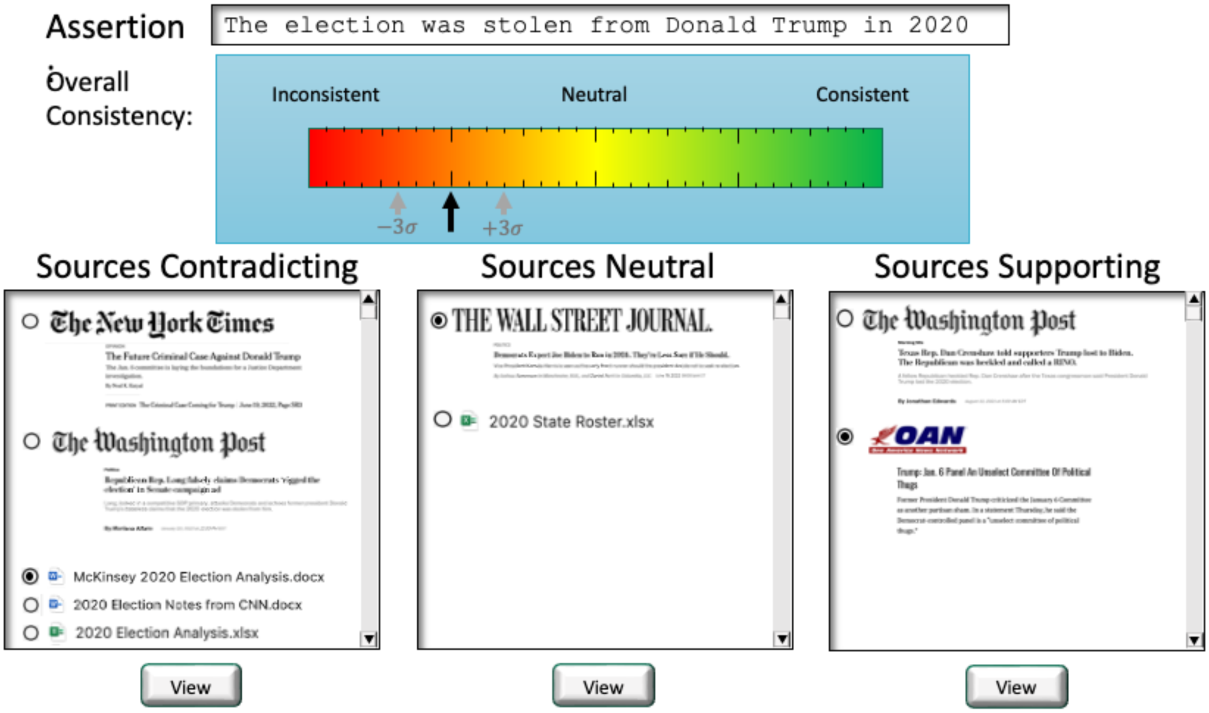


Figure 7. Flow Chart for a Text Analysis Toolset

One useful feature that could be implemented is shown in the last step Figure 7 (evidence-based analysis). In this step, a user enters a brief text-based statement. DNA's output is a list of sources that echo or reinforce the statement, a list of sources that disagree with or contradict the statement, and the specific texts in each that are interpreted as indicating reinforcement or contradiction. An example would be to input the phrase "Donald Trump won the 2020 election," whereby DNA would return a list of the news articles supporting the statement and those contradicting it, as shown below.



The election was stolen from Donald Trump in 2020

Sources Contradicting

○ The New York Times
The Future Criminal Case Against Donald Trump

○ The Washington Post
Republican Rep. Long falsely claims Democrats 'rigged the election' in Senate campaign ad

● McKinsey 2020 Election Analysis.docx
○ 2020 Election Notes from CNN.docx
○ 2020 Election Analysis.xlsx

Sources Neutral

● THE WALL STREET JOURNAL.
Democrats Expect Joe Biden to Run in 2024. They're Less Sure if He Should.

○ 2020 State Roster.xlsx

Sources Supporting

○ The Washington Post
Texas Rep. Dan Crenshaw told supporters Trump lost to Biden. The Republican was heckled and called a RINO.

● OAN
Trump: Jan. 6 Panel An Unselect Committee Of Political Thugs
Former President Donald Trump criticized the January 6 Committee as another partisan sham. In a statement Thursday, he said the Democrat-controlled panel is a "unselect committee of political thugs."

Figure 8. Evidence-Based Analysis

6.2 Illustrating to Readers How a Text May Be Biased

In order to inform a reader as to how their news source may be biasing information, an interface comparing a specific article from their source (such as the New York Times or Fox News in the United States) with a centrist publication could be made available. This interface would utilize DNA to perform a detailed analysis of the article identified by the user (and then scraped from a web site or accessed via a news API), and then compare the results with an article on the same subject from a different, more "centrist" source.

Information (similar to and building on the details discussed in Section 4) could be presented to illustrate how the text may be biased as regards the topics, events, people and locations that are discussed (and how they are described), the sources and quotations that are referenced, and more.

6.3 Other Application Areas

DNA has the potential to enable a broad range of applications in a variety of sectors. A few examples are:

- Analysis and comparison of text-based, ML data sets to understand differences in training versus test sets (if ML models are not performing adequately on a test set), or training versus application sets (to understand if an ML model is applicable to the data set)
- Analysis and comparison of corporate documentation to highlight inconsistencies
- Historical analysis (aggregating and comparing family histories, historical records, etc.)

- Medical transcript analysis (of both patient and provider narratives)
- Support for government and NGO aid and research into social and economic conditions (using narratives gathered from the local population and other sources)
- Law and law enforcement (comparison of dispositions and other narrative forms)
- Political discourse analysis (interviews, speeches, platforms, etc.)

7.0 CONCLUSIONS

DNA's goal is to enable users and third-party applications to perform detailed comparisons and analyses of texts. Possible usage scenarios involve understanding where texts align and diverge, what events and conditions are mentioned (or omitted), what words are used to elicit emotions in readers, and much more. In this way, DNA can be used to analyze and illustrate how, for example, a news article or blog post is biased, or if there are certain themes that are consistent across a set of narratives and articles. Regarding themes in narratives, DNA can be used to find evidence supporting or refuting hypotheses, or regarding which combination of circumstances, actions and events are correlated with "success" or "failure" in a situation.

DNA's approach is built on the combined application of semantic, linguistic, ontological and machine-learning technologies. Although work on DNA is still in an early research stage, its architecture and implementation are available as open-source for review and collaboration.

8.0 REFERENCES

[1] Almeida, J., Falbo, R. & Guizzardi, G. (2019). Events as Entities in Ontology-Driven Conceptual Modelling. In: Laender, A., Pernici, B., Lim, EP., de Oliveira, J. (editors). Conceptual Modeling. ER 2019. Lecture Notes in Computer Science, vol 11788. Springer, Cham. doi: 10.1007/978-3-030-33223-5_39. Retrieved from https://link.springer.com/chapter/10.1007/978-3-030-33223-5_39.

[2] Bratanic, T. (2021). From Text to Knowledge: The Information Extraction Pipeline. Blog post. Retrieved from https://towardsdatascience.com/from-text-to-knowledge-the-information-extraction-pipeline-b65e7e30273e.

[3] Chiusano, F. (2022). Building a Knowledge Base from Texts: A Full Practical Example. Blog post. Retrieved from https://medium.com/nlplanet/building-a-knowledge-base-from-texts-a-full-practical-example-8dbbffb912fa.

[4] Gao, Y., Song, S., Zhu, X., Wang, J., Lian, X. & Zou, L. (2018). Matching Heterogeneous Event Data. IEEE Transactions on Knowledge and Data Engineering. Volume 30, Number 11, pp. 2157-2170. doi: 10.1145/2588555.2588570. Retrieved from https://ieeexplore.ieee.org/document/8315460.

[5] Hamilton, K. (2021). Towards an Ontology for Propaganda Detection in News Articles. In: The Semantic Web: ESWC 2021 Satellite Events (ESWC 2021). Lecture Notes in Computer Science, vol 12739. Springer, Cham. doi: 10.1007/978-3-030-80418-3_35. Retrieved from https://link.springer.com/chapter/10.1007/978-3-030-80418-3_35.

[6] Honnibal, M. & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Product web site, https://spacy.io/.

[7] Hu, B., Wang, J. & Zhou, Y. (2009). Ontology Design for Online News Analysis. 2009 WRI Global Congress on Intelligent Systems, pp. 202-206. doi: 10.1109/GCIS.2009. Retrieved from https://ieeexplore.ieee.org/document/5209304.

[8] Ilievski, F., Oltramari, A., Ma, K., Zhang, B., McGuinness, D. & Szekely, P. (2021). Dimensions of commonsense knowledge. Knowledge-Based Systems. Volume 229. doi: 10.48550/arvix.2101.04640. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0950705121006092.

[9] Linguistic Data Consortium (2005). ACE (Automatic Content Extraction) English Annotation Guidelines for Events. Version 5.4.3. Retrieved from https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf.

[10] Liu, K., Chen, Y., Liu, J., Zou, X. & Zhao, J. (2020). Extracting Events and Their Relations from Text: A Survey on Recent Research Progress and Challenges. AI Open. Volume 1, pp. 22-39. doi: 10.1061/j.aiopen.2021.02.004. Retrieved from https://www.sciencedirect.com/science/article/pii/S266665102100005X.

[11] Marcus, G. & Murphy, E. (2022). Three Ideas from Linguistics that Everyone in AI Should Know. Blog post. Retrieved from https://garymarcus.substack.com/p/three-ideas-from-linguistics-that.

[12] McShane, M. (2017). Natural Language Understanding (NLU, not NLP) in Cognitive Systems. AI Magazine. Volume 38, Issue 4, pp. 43-56. doi: 10.1609/aimag.v38i4.2745. Retrieved from https://ojs.aaai.org/index.php/aimagazine/article/view/2745.

[13] McShane, M. & Nirenburg, S. (2021). Linguistics for the Age of AI. MIT Press. ISBN: 9780262045582. Retrieved from https://mitpress.mit.edu/9780262045582/linguistics-for-the-age-of-ai/.

[14] Mitamura, T., Liu, Z. & Hovy, E. (2017). Events Detection, Coreference and Sequencing: What's next? Overview of the TAC KBP 2017 Event Track. TAC 2017. Retrieved from https://hunterhector.github.io/files/papers/Mitamura,_Liu,_Hovy_-_2018_-_TAC_2017.pdf.

[15] Mostafazadeh, N., Kalyanpur, A., Moon, L., Buchanan, D., Berkowitz, L., Or, B. & Chu-Carroll, J. (2020). GLUCOSE: Generalized and Contextualized Story Explanations. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4569-4586. doi: 10.18653/v1/2020.emnlp-main.370. Retrieved from http://aclanthology.lst.uni-saarland.de/2020.emnlp-main.370.pdf.

[16] Noy, N. & McGuinness, D. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880. Retrieved from http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf.

[17] OntoInsights, LLC (2022). Deep Narrative Analysis RESTful APIs v1.0.2. Online documentation. Retrieved from https://ontoinsights.github.io/dna-swagger/.

[18] Princeton University. About WordNet. https://wordnet.princeton.edu/. Princeton University. 2010.

[19] Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A. & Choi, Y. (2019). ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33, Issue 01, pp. 3027-3035. doi: 10.1609/aaai.v33i01.33013027. Retrieved from https://ojs.aaai.org//index.php/AAAI/article/view/4160.

[20] Scherp, A., Franz, T., Saathoff, C. & Staab, S. (2010). A core ontology on events for representing occurrences in the real world. Multimedia Tools and Applications. Volume 58, pp. 293–331 (2012). doi: 10.1007/s11042-010-0667-z. Retrieved from https://link.springer.com/article/10.1007/s11042-010-0667-z.

[21] Schrodt, P. (2012). CAMEO Conflict and Mediation Event Observations. Event and Actor Codebook. Retrieved from http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf.

[22] Shortwave, NPR Whyy. The Science Behind Storytelling podcast (19 August 2020). Retrieved from https://www.npr.org/2020/08/18/903545336/the-science-behind-storytelling.

[23] Truswell, R. (editor, 2019). The Oxford Handbook of Event Structure. Oxford University Press. doi: 10.1093/oxfordhb/9780199685318.001.0001.

[24] University of Michigan (2022). “Fake News,” Lies and Propaganda: How to Sort Fact from Fiction. Online research guide. Retrieved from https://guides.lib.umich.edu/fakenews.

[25] Zeng, K., Li, C., Hou, L., Li, J. & Feng, L. (2021). A comprehensive survey of entity alignment for knowledge graphs. AI Open, Volume 2, pp. 1-13. doi: 10.1016/j.aiopen.2021.02.002. Retrieved from https://www.sciencedirect.com/sciendirect.com/science/article/pii/S2666651021000036.