The Battelle Sematrix™ Advantage

Introduction

Each year, researchers add hundreds of thousands of articles to the peer-reviewed literature. Findings from most studies will take 17 years or more to translate into clinical practice (1). Few scientists expect their individual works to result in practice changes, but most expect that their work will contribute to the knowledgebase that informs practice, as well as informing future research and policymaking. Evidence-based practice (EBP) changes require a time-consuming and labor-intensive process of selecting, analyzing and synthesizing the evidence necessary to address critical clinical questions. The results of this process intend to help busy clinicians stay abreast of current evidence. The limited availability of financial and human resources required for evidence analysis and synthesis delays translation of evidence to clinicians (and ultimately patients). The lag in evidence—and, thus, knowledge—translation has many causes, but at Battelle we believe that the cost and burden of the evidence selection, analysis and synthesis process causes investigators to limit their reviews to those publications that reflect only major facets of a disease or a limited population. These circumstances make exhaustive, manual review a rarity, leaving scientists, clinicians and patients without all of the necessary evidence, because investigators are forced to generalize from imperfect, incomplete and inadequate evidence.

To overcome these limitations, innovators at Battelle have developed Sematrix™, a program that supports automation of the reviewing tasks of document collection, filtering, comparison and—to some extent—synthesis of the knowledge contained in an entire corpus of documents. Sematrix is a sequence of linked processes that add value to unstructured scientific or technical information usually found in published literature. It does this by extracting axiomatic knowledge from text and representing it graphically in a form that supports reasoning.

Early in its development, the Battelle team demonstrated the value of using robust natural language processing (NLP) to select knowledge entities from predefined, ontology-based categories in the sample domain of antibiotic resistance. Before the development of any knowledge-using application, which generally involves semantic query, Sematrix constructs a knowledgebase in the form of axioms or “triples” using a formal structure known as the Resource Description Framework (RDF). RDF is a commonly used format for representing linked data (e.g., “[the intervention] handwashing [the outcome] reduces person-to-person spread of [the target] staphylococcus”), but RDF alone supports only limited reasoning through subsumption relationships. To maximize Sematrix’s capabilities, Battelle added an “inference layer” of processing rules to expand the reasoning power of RDF data. With hundreds of measures in need of development, CMS and Battelle partnered to find a better way.

The next phase focused on defining and understanding the potential uses of that extracted knowledge and its representation scheme to support advanced analytics. Using OWL 2 with RDF triples, Sematrix overcomes many of the established limitations in the current NLP and knowledge representation. This is because Battelle anticipated potential problems while focusing on useful applications. Most problems were foreseeable and had direct solutions. For others, we continue to develop new, advanced language- and concept-processing capabilities.

These issues include:

- Automatic concept formation
- Aboutness thresholds
- Determining degrees of similarity
- Verifying uniqueness
- Validating assertions
- Validating logic
- Normalizing visualization
- Interpreting network graph features
- Lowering the costs of extraction
- Foreign language knowledge representation
- Coping with discourse modalities.

What Sematrix Does

Sematrix represents the knowledge from scientific and technical documents in a way that allows investigators to identify and map similar research, identify gaps in research conducted within a given field, or map relationships among certain authors or groups. This capability depends on a granular extraction of scientific information from unstructured text using the RDF output format (i.e. axiomatic "triples"). This granularity also permits continuous enlargement of the dataset, thus enhancing the set of possible analytics that we can perform on the extracted data, such as data manipulation and visualization.

Conversion of textual content into a large collection of fine-grained axiomatic triples is extremely useful. It allows the system to automatically summarize a paper's content in a scientifically valid way (e.g. with the knowledge that an intervention influences an outcome–a distinction that is difficult, though not impossible, to make using a purely statistical approach to document categorization). More importantly, this conversion identifies the patterns of research activity and findings that a paper intends to describe, which allows contemplation of several applications, among them comparison of research techniques in a given domain or across several domains. This empowers the system to establish patterns of research attributable to a given researcher or research team.

Our initial deployment of this program supports specific, focused reviews, but we are investing in development that will facilitate experts' and clinicians' rapid collection of evidential documents for cases or populations that fall between the cracks of ordinary environmental scans. Currently, Sematrix supports investigators' engagement in comprehensive, efficient, and cost-effective literature reviews within the very limited literature in which we have trained it. This will expand as we train the system to "understand" more textual inputs. Ultimately, we will develop Sematrix to automate most of the time-consuming processes needed to conduct environmental scans, especially the processes of selection, analysis and synthesis of inputs.

Many recent text-analytics systems designs rely on ontologies—lexical category schemas with classes of hierarchically related entities (taxonomies) and horizontally related or network patterns that give semantic values to relationships among the classes. Researchers have recently applied high-level semantic categorization to comprehensive medical terminology systems like the Unified Medical Language System (UMLS). The UMLS articulates five major categories of non-hierarchical relationships: physical, spatial, temporal, functional and conceptual. Even in specific domains, these limited inter-entity relationship sets cannot adequately represent the nuances of natural language. Sematrix is a linked set of computational processes, including NLP, that convert textual narrative into a form that can be interrogated, analyzed and understood by predicate logic and statistics. Basic ontologies provide the "aboutness" of a given document by assigning extracted text segments to ontological categories, but the lack of first-order logical characteristics to specify relationships between any two classes limits deep reasoning from text (especially across documents). Sematrix addresses this deficiency in two ways:

1. Sematrix automates the production of provisional ontologies from a topical corpus (e.g., respiratory failure) and
2. Sematrix describes in detail the logic features of axiomatic properties (e.g., domain and range specification, directionality) required for inferential (deductive) reasoning about the claims in the text.

Representation of sentential meaning using the Sematrix process provides an advantage over other text-analytics systems that rely primarily on statistical term clustering because Sematrix efficiently captures the axiomatic knowledge pertaining to a certain domain in context, as represented by some set of traversable properties or relationships to other axioms. English writers often use several levels of subordination or conjunction to express complex ideas or concepts. Sematrix translates these into nested or linked sets of predicates. We organize these for analytic purposes into the commonly used PICO (Population,

Intervention, Comparison, Outcome—and sometimes Setting). This translation capability addresses Huang's complaint that some semantic classes can be tied back to multiple PICO slots. [Treatment and Drug] [Disease], and [Symptom] are examples of semantic classes that take different roles for different question types. Significant context is required for accurate mapping (2).

Sematrix does not intend to produce environmental scans per se, but to locate, extract and organize the best available knowledge. Thus, beyond categorizing text portions according to a predicate-logic-based ontology that facilitates inferential reasoning and the knowledge-pattern matching, the system should also answer questions by isolating sentences, sentence fragments, passages and paragraphs that correctly interpret and address the research questions. Battelle recently demonstrated that Sematrix NLP can examine a corpus of grant proposals to identify the best answers, both across the corpus and within a given proposal, to questions like, "What are the desired outcomes?" or "What are the primary barriers and challenges?" Rather than reconstituting the answers to these questions by interrogating the low-level axioms in the knowledgebase, it is simpler to parse the input documents to locate natural language sentences in context that could contribute to the answer. This approach gives the ability to cope with answers that are non-contiguous and/or partial. Using a correspondence ranking algorithm, users could set thresholds to exclude complete but poor answers.

Summary

Sematrix is a complex linked set of computational processes, including NLP, that convert textual narrative into a form that can be interrogated and analyzed, and to some extent understood, by means of both predicate logic and statistical analysis. Properly deployed and applied, Sematrix will assist in the more onerous tasks of environmental scanning, including document retrieval (selection), knowledge categorization/representation and development of research questions. Sematrix can support two similar, but distinct, cognitive processes needed in SR, namely knowledge-pattern matching and question answering. With semantic and cognitive processing, the next step is the development of knowledge models that can represent both the objects in a knowledgebase and the query that encodes the user's true information/knowledge needs. Early methods for building weighted document profiles relied solely on term frequency and probability to represent a document (for example, see Salton and Yang [3]). We can now evaluate document similarity based on correspondence with a highly granular semantic network structure. Using PICO to articulate clinical questions, combined with the ability to represent the knowledge within a given document as a graph (because an ontology is also a graph), we can calculate the correspondence between the query and the document knowledge set—a more robust method of indicating relevance than simple frequency. This allows investigators to set virtually any criteria to weight and sort the knowledge patterns found in a large number of research articles. Battelle innovators believe that this tool will broadly influence the direction and comprehensiveness of scientific inquiry for the next decade.

Reference List

(1) Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med 2011; 104 C(12):510-520.

(2) Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. American Medical Informatics Association. Annu Symp Proc 2006; 2006:359-363.

(3) Salton G, Yang CS. On the Specification of term values in automatic indexing. J Doc 1973; 29(4):351-372.

Contributors

Darryl W. Roberts, PhD, MS, RN
Bruce A. McClelland, PhD, MFA
Jeremy Bellay, PhD
Nancy J. McMillan, PhD

Every day, the people of Battelle apply science and technology to solving what matters most. At major technology centers and national laboratories around the world, Battelle conducts research and development, designs and manufactures products, and delivers critical services for government and commercial customers. Headquartered in Columbus, Ohio since its founding in 1929, Battelle serves the national security, health and life sciences, and energy and environmental industries. For more information, visit www.battelle.org.

800.201.2011 | **solutions@battelle.org** | **www.battelle.org**

Battelle and its logos are registered trademarks of Battelle Memorial Institute. © Battelle Memorial Institute 2017. All Rights Reserved.

ID 543 08/17

BATTELLE

It can be done