Understanding and Exploring Competitive Technical Data from Large Repositories of Unstructured Text

James J. Nolan
Decisive Analytics Corporation
Arlington, VA, USA
jim.nolan@dac.us

Mark Stevens
Decisive Analytics Corporation
Jeffersonville, IN, USA
mark.stevens@dac.us

Peter David
Decisive Analytics Corporation
Arlington, VA, USA
peter.david@dac.us

ABSTRACT

We present an approach to automatically processing open source unstructured data to extract relevant technical information. The approach is tailored towards technology monitoring, and specifically to prevent “technical surprise” - when a competitor or adversary develops and deploys an unexpected technology. Our approach takes advantage of Natural Language Processing, Entity Extraction, and Visual Document Processing. We provide an intuitive interface that allows users to easily interact with the Machine Learning system.

Author Keywords

Technology tracking, natural language processing, semantic reasoning, entity extraction, relationship extraction, directed exploration.

ACM Reference format:

Nolan, James, Stevens, Mark, and David, Peter. 2019. Understanding and Exploring Competitive Technical Data from Large Repositories of Unstructured Text. In Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019, 5 pages.

INTRODUCTION

Technology across the global landscape is changing at a record pace. Keeping track of or discovering this information in a timely fashion remains a difficult challenge.

There are many use cases where it is important to prevent “technical surprise”, when a competitor or adversary develops and deploys an unexpected technology. For example, consider smart phone technology. Smart phone manufacturers, such as Apple or Samsung, would like to know immediately when one of their competitors develops a chip that outperforms previous generations or develops a new glass with greater drop resistance. Consider military adversaries as another example. Military leaders need to know when adversaries develop new planes or weapons that can fly higher or further than before.

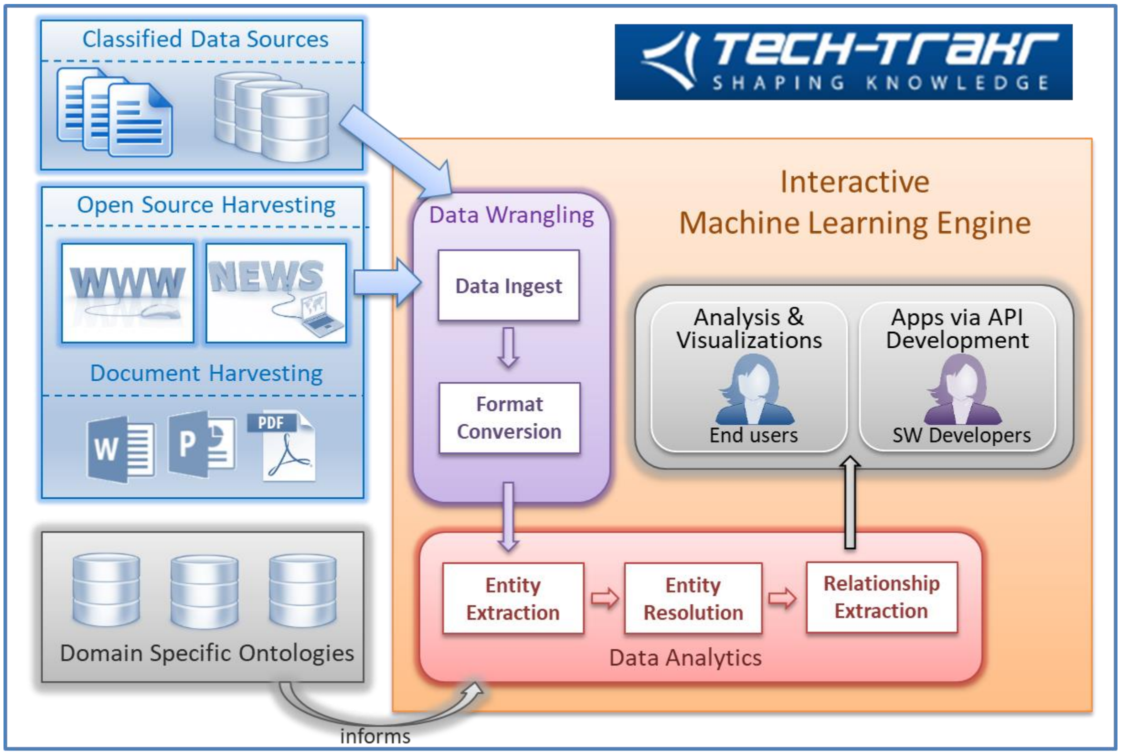


Figure 1 - An Overview of the Tech-Trakr approach.

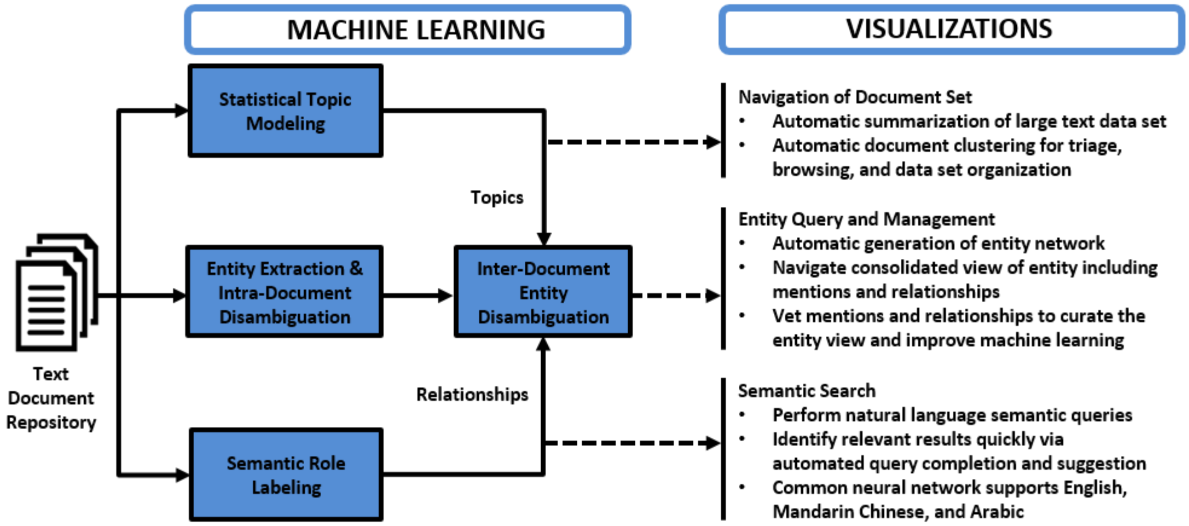

MACHINE LEARNING
VISUALIZATIONS
Statistical Topic Modeling
Entity Extraction & Intra-Document Disambiguation
Inter-Document Entity Disambiguation
Semantic Role Labeling
Text Document Repository
Topics
Relationships
Navigation of Document Set
• Automatic summarization of large text data set
• Automatic document clustering for triage, browsing, and data set organization
Entity Query and Management
• Automatic generation of entity network
• Navigate consolidated view of entity including mentions and relationships
• Vet mentions and relationships to curate the entity view and improve machine learning
Semantic Search
• Perform natural language semantic queries
• Identify relevant results quickly via automated query completion and suggestion
• Common neural network supports English, Mandarin Chinese, and Arabic

Figure 2 - An Overview of the NLP Capabilities Utilized by Tech-Trakr

The challenge to tracking technology to avoid technical surprise is due in part to the fact that manufacturers attempt to guard this information, choosing not to publish it for fear of losing their competitive advantage. However, in spite of this guarding of information, the data does frequently end up in the open source domain through industry publications, journal/conference proceedings, news sources, or press releases. Given that this information does make it out into the unstructured wild, finding, extracting, and tagging the information, and ultimately putting it into a format and structure that can be used for competitive analysis is an enormous challenge.

To address this problem, we developed a tool called Tech-Trakr,[1] which encapsulates a suite of Natural Language Processing (NLP) and Machine Learning (ML) capabilities to perform automated extraction and support directed exploration of competitive technical data from unstructured text. In this paper, we (1) provide an overview of the Tech-Trakr system and underlying technologies, and (2) present a case study to exemplify how Tech-Trakr supports directed exploration for understanding a particular technology.

TECH-TRAKR OVERVIEW

An overview of the Tech-Trakr tool is illustrated in Figure 1. Working from left to right, Tech-Trakr harvests data from the web using results from commercial search engines, as well as focusing on specific sites of interest, news sources, and classified document collections. Ingested data is formatted and normalized for processing to provide clean data to the downstream algorithms. Tech-Trakr automatically identifies specific entities, resolves them to remove ambiguity, extracts metadata and values that describe those entities, and identifies relationships between them. These analytics are informed by domain-specific ontologies that encode expertise for optimization. The entity and relationship information that is output by the analytics is then stored in a database and can be accessed via an API or through custom visualizations.

Figure 2 provides more detail on the NLP capabilities and visualizations embedded within Tech-Trakr. For the purposes of this paper, we focus on two key enablers for updating technology databases from unstructured data: Entity Extraction and Relationship Discovery and how they provide actionable, database-quality information from unstructured data.

BACKGROUND

Tech-Trakr relies on four primary NLP capabilities that automatically process and provide insights into large unstructured text repositories: Dealing with diverse data, Statistical Topic Modeling (STM), Semantic Role Labeling (SRL), and Entity Extraction and Disambiguation.

Dealing with Diverse Unstructured Data Sets

Tech-Trakr provides the ability to collect, parse, and extract relevant information from diverse and unstructured data sets. Collection from such a large number of sources produces data with extreme variety of file formats, document organization, page layout, text style, and content. This extreme document variety makes it difficult to perform even simple tasks such as understanding the content and how it impacts analysis. A machine learning capability that can automate skills that analysts perform well, while also scaling up to handle data velocity is critically needed. Tasks such as extracting document titles, authorship information, security classification, or top-level headings are difficult and time-consuming processes.

[1] http://techtrakr.dac.us/techtrakr-ui/#/about

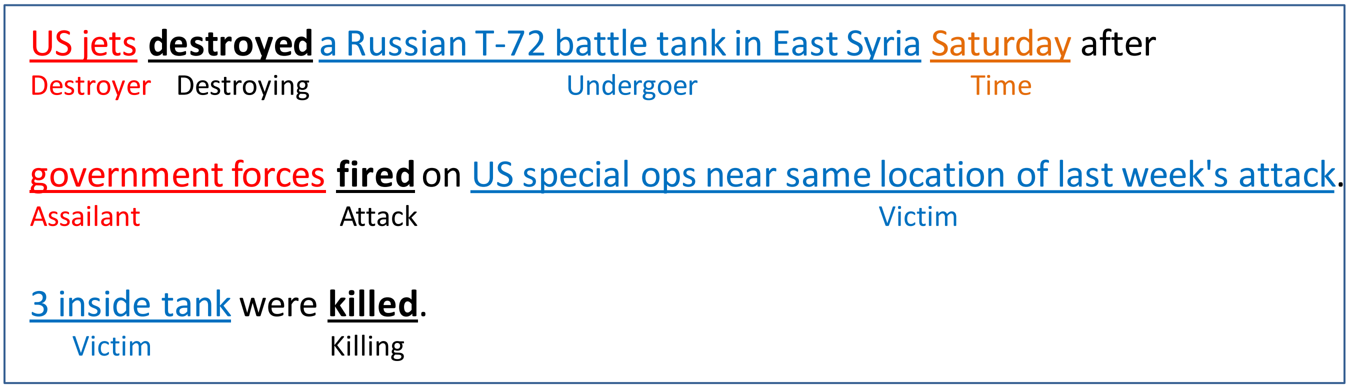

US jets destroyed a Russian T-72 battle tank in East Syria Saturday after
Destroyer
Destroying
Undergoer
Time
government forces fired on US special ops near same location of last week's attack.
Assailant
Attack
Victim
3 inside tank were killed.
Victim
Killing

Figure 3 – Semantic Role Labeling Example

One major cause of this difficulty is the proliferation of metadata-less formats such as PDF files of scanned documents and text data formatted solely through improvised or informally defined typographic conventions. These data lack an underlying, machine-readable explanation of how text formatting and style represent document structure and metadata.

Tech-Trakr provides two foundational machine learning capabilities for dealing with these issues:

Data Harvesting

Tech-Trakr acquires content through both ingest of internally-maintained collections of documents and by initiating web searches for online content. Tech-Trakr's ingest pipeline is designed to process data with extreme heterogeneity in terms of file formats, document organization, page layout, text style, and content. The on-line retrieval function runs periodically, retrieving new content when it is available online.

Visual Document Processing

Tech-Trakr includes a Visual Document Processing (VDP) capability that uses visual analysis of documents to infer the communicative intent of the author and to recover document structure and metadata. Our algorithm identifies the components of documents such as titles, headings, and body content, based on their appearance. Our algorithm is entirely format-agnostic; it does not rely on document mark-up or metadata to identify the structural components of a document. Instead, it operates on an image of a document and can learn from any document type, including scanned images.

Statistical Topic Modeling

Statistical topic modeling [1] discovers topics and clusters documents to support rapid exploration of data.

Semantic Role Labeling (SRL)

SRL extracts meaning from sentences by identifying and labeling semantic predicates and arguments. SRL is an NLP process that maps the words and phrases in unstructured text to a formal model of text meaning. In our prior work, we developed an SRL capability that analyzes the semantics of whole sentences, identifies the fundamental concepts, called *frames* [2], that are discussed, maps words and phrases from the text to the roles that are related to these concepts, and ultimately updates structured databases. Our SRL capability has been used to perform analysis of open source data [3], extract rich social network structures from unstructured text [4], and accurately extract entities from unstructured data and map information about them to structured databases. An example of the output of our SRL capability is shown in Figure 3.

Figure 3 illustrates how SRL maps words and phrases in text to a structured model of meaning. Our event-oriented model of meaning defines hundreds of event types, such as Destroying, Attack, and Killing. The SRL algorithm determined that the words *destroyed*, *fired*, and *killed* in the sample text evoke these event types. Other phrases in the text, such as *US jets* were mapped to event-specific roles. This SRL capability is at the core of our Tech-Trakr product. Tech-Trakr uses SRL to find the relationships between products, manufacturers, and other entities. Tech-Trakr stores the extracted information in a database, allowing downstream analytics to retrieve information about events, relationships between entities, and attributes of those entities.

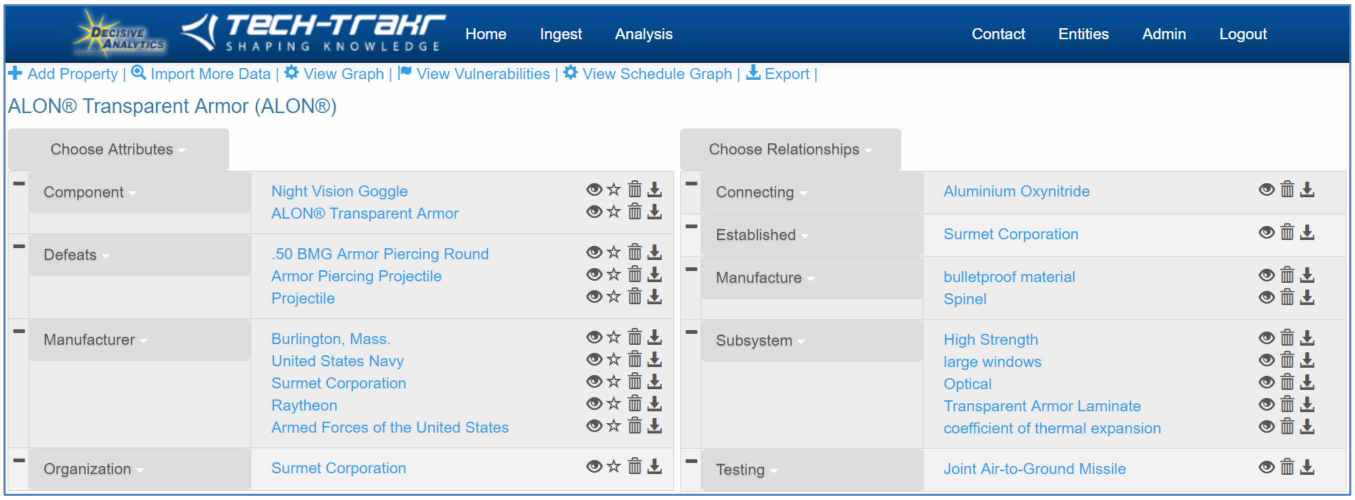
Entity Extraction and Disambiguation

Entity extraction and disambiguation provide a consolidated view of an entity across the entire text data set. [5]

CASE STUDY: TRANSPARENT ARMOR

As a working example of Tech-Trakr's capabilities, consider an analyst tasked with assessing industry's Transparent Armor[2] capabilities. An analyst can perform an open-source search to quickly discover that compounds Aluminum Oxynitride (ALON) and Aluminum Oxide (Sapphire) are critical components for transparent armor. While discovering

[2] Transparent Armor is a type of bullet proof glass that can be worn to prevent injury



ALON® Transparent Armor (ALON®)

Attribute	Values
Component	Night Vision Goggle ALON® Transparent Armor
Defeats	.50 BMG Armor Piercing Round Armor Piercing Projectile Projectile
Manufacturer	Burlington, Mass. United States Navy Surmet Corporation Raytheon Armed Forces of the United States
Organization	Surmet Corporation

Relationship	Values
Connecting	Aluminium Oxynitride
Established	Surmet Corporation
Manufacture	bulletproof material Spinel
Subsystem	High Strength large windows Optical Transparent Armor Laminate coefficient of thermal expansion
Testing	Joint Air-to-Ground Missile

Figure 4 - Extracted categories and their Values for Transparent Armor

the importance of these components may be a simple task, it is significantly more difficult to determine all of the important characteristics of these materials and present them in a meaningful way for sharing with other analysts. Additionally, consider that ALON and Sapphire are two of a potentially large set of Transparent Armor materials of interest to an analyst. The challenge is accurately extracting every relevant material and guaranteeing that this information is up-to-date and accurate. In other words, the challenge is extracting all relevant information for all technologies of interest, at scale, as it becomes available.

Now let us consider specifically the Transparent Armor use case and walk through how Tech-Trakr automatically populates a database with information that can be used to generate a detailed profile about this technology.

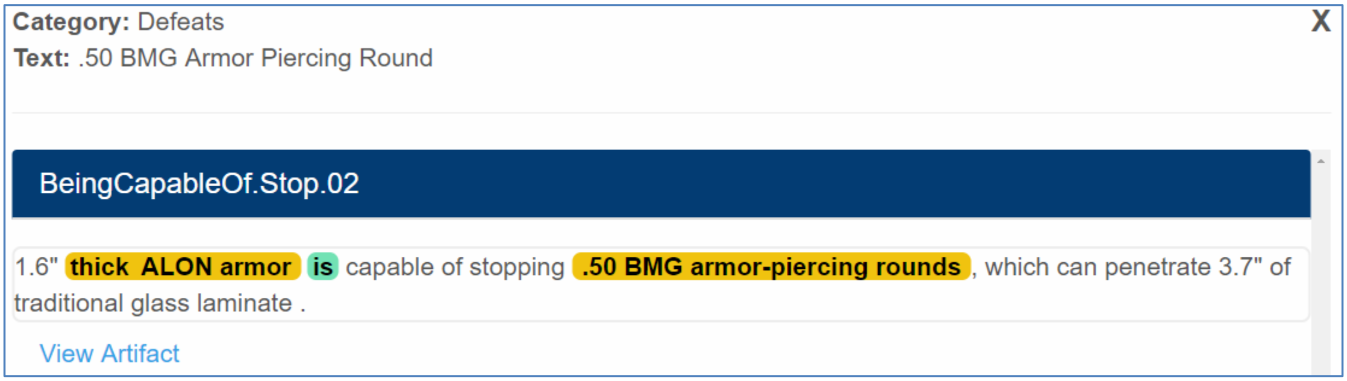
Technology Profile

In Figure 4, we show Tech-Trakr's automatically generated profile of the Transparent Armor technology which includes extracted attributes and relationships related to the technology. These extracted attributes and relationships are those that Tech-Trakr has identified as important to most accurately capture the essence of Transparent Armor. Within the Component section, for example, Tech-Trakr has automatically linked ALON to Night Vision Goggles and Joint Air-to-Ground Missiles. Additionally, Tech-Trakr extracted the chemical composition (Aluminum Oxynitride), the manufacturer (Surmet Corporation), the manufacturing location (Burlington, Mass.), and the claim that it can defeat a .50 BMG Armor Piercing Round. We are displaying only a small subset of the over 100 categories of information that have been discovered about this ballistic glass that comprises Transparent Armor.

The complete Tech-Trakr profile for Transparent Armor consists of over 200 additional characteristics, correlations, and relationships, and was extracted from less than 500 articles discovered via open source harvesting capability.

Directed Source Exploration

The Tech-Trakr profile shown in Figure 4 is interactive and allows the analyst to drill into the source material from which the relevant information was extracted. The attribute values highlighted in blue are named entity hyperlinks that navigate to more information about that concept in the form of its own entity profile. The icons to the right of each attribute value provide advanced user options and information. Clicking the icon that resembles an eye navigates to an annotated view of the source data from which the attribute value was extracted. For example, in Figure 5, within the Defeats section of the Transparent Armor profile, there is an entry for the .50 BMG



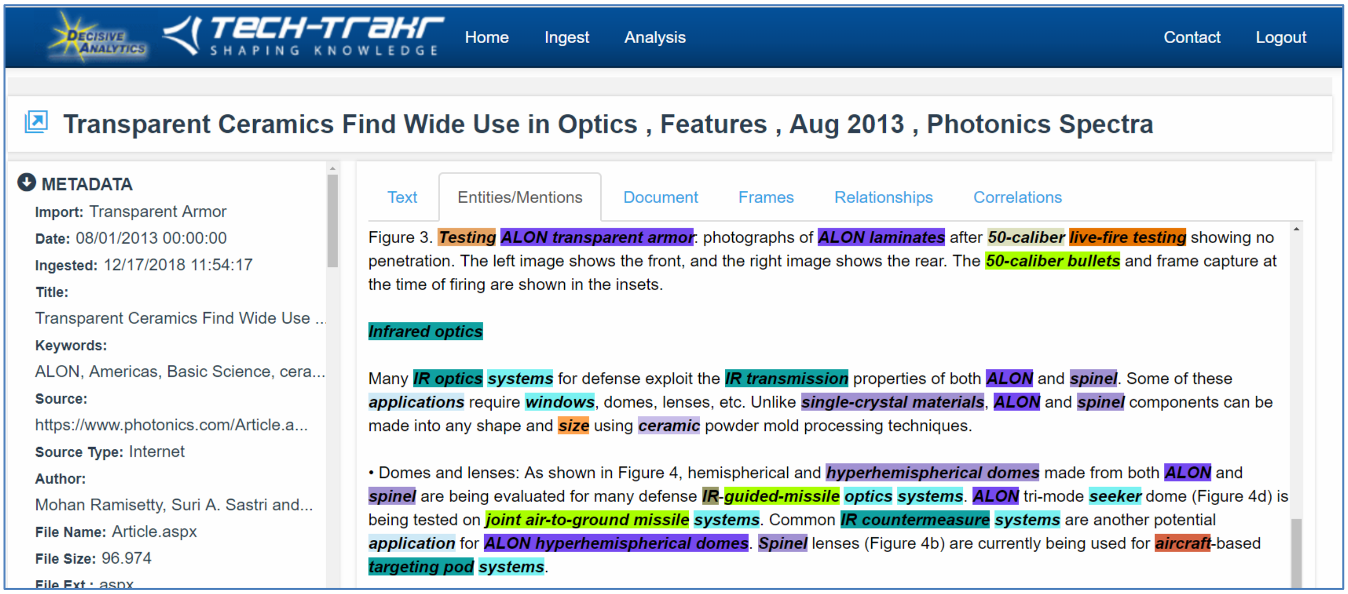
Category: Defeats
Text: .50 BMG Armor Piercing Round

BeingCapableOf.Stop.02

1.6" thick ALON armor is capable of stopping .50 BMG armor-piercing rounds, which can penetrate 3.7" of traditional glass laminate .

View Artifact

Figure 5 - Exploring the "Defeats" category to Determine a Type of Munition incapable of penetrating Transparent Armor



Transparent Ceramics Find Wide Use in Optics , Features , Aug 2013 , Photonics Spectra

METADATA

Import: Transparent Armor
Date: 08/01/2013 00:00:00
Ingested: 12/17/2018 11:54:17
Title:
Transparent Ceramics Find Wide Use ...
Keywords:
ALON, Americas, Basic Science, cera...
Source:
https://www.photonics.com/Article.a...
Source Type: Internet
Author:
Mohan Ramisetty, Suri A. Sastri and...
File Name: Article.aspx
File Size: 96.974
File Ext.: aspx

Figure 3. Testing ALON transparent armor: photographs of ALON laminates after 50-caliber live-fire testing showing no penetration. The left image shows the front, and the right image shows the rear. The 50-caliber bullets and frame capture at the time of firing are shown in the insets.

Infrared optics

Many IR optics systems for defense exploit the IR transmission properties of both ALON and spinel. Some of these applications require windows, domes, lenses, etc. Unlike single-crystal materials, ALON and spinel components can be made into any shape and size using ceramic powder mold processing techniques.

• Domes and lenses: As shown in Figure 4, hemispherical and hyperhemispherical domes made from both ALON and spinel are being evaluated for many defense IR-guided-missile optics systems. ALON tri-mode seeker dome (Figure 4d) is being tested on joint air-to-ground missile systems. Common IR countermeasure systems are another potential application for ALON hyperhemispherical domes. Spinel lenses (Figure 4b) are currently being used for aircraft-based targeting pod systems.

Figure 6 - The Tech-Trakr Approach enables exploration down to the original source document

Armor-Piercing Round. An analyst interested in understanding how the system determined that Transparent Armor has a "Defeats" relationship with this caliber of ammunition can click the eye icon on that row of the profile, which navigates to the annotated source data view shown in Figure . This view displays the source sentence and the name of the semantic frame from which the attribute or relationship was identified. Additionally, the source sentence is annotated with the recognized roles of the semantic frame including the verb that evoked the frame. The "View Artifact" link located beneath the source sentence navigates to the original annotated document complete with metadata outlining the originating source as well as the collection and creation date of the document, as shown in Figure 6.

CONCLUSION

In this paper we present a tool called Tech-Trakr that automatically extracts and provides analysts with an overview and directed exploration of technical data from unstructured text. This tool is based on NLP techniques, including SRL and entity extraction and disambiguation to automatically extract and organize information relevant to various technologies. Tech-Trakr produces technology profiles containing relevant information, such as chemical composition, capabilities, strength, durability, and alternate applications. Analysts interact with the profiles to explore the relevant source data and gain additional understanding of the technology. We demonstrate the Tech-Trakr capability using a specific use case of understanding Transparent Armor technology.

REFERENCES

1. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." *The Journal of Machine Learning Research* 3 (2003): 993–1022.
2. Ruppenhofer, Josef, Michael Ellsworth, Miriam RL Petruck, Christopher R. Johnson, and Jan Scheffczyk. *FrameNet II: Extended Theory and Practice*, 2010. http://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf.
3. Kase, Sue E. "Accelerating Exploitation of Low-Grade Intelligence through Semantic Text Processing of Social Media." DTIC Document, 2013. http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA587022.
4. Davenport, Jack H., and James J. Nolan. "Social Network Analysis Realization and Exploitation." Baltimore, MD, 2015.
5. Ward, Kevin, and Jack Davenport. "Human-Machine Interaction to Disambiguate Entities in Unstructured Text and Structured Datasets." Anaheim, CA, 2017.