Semantic Text Analysis to Identify Risks in Architecture Design and Recurring Problems in Report Databases

Jane T. Malin, *Johnson Space Center*
David R. Throop, *The Boeing Company*

Mozilla Firefox

File Edit View Go Bookmarks Tools Help

Equipment	Failure	Examples	Sentence
BEARING	1.1.1.1. Uneven [surface]: uneven, flawed, ridged, rough, gritty, non-uniform, bumpy, irregular, marred		
	ROUGHEN	CA453	NEEDLE BEARINGS AND BROKEN BEARING PARTS WERE ROUGHENED
	GRIT	CD205	FOUND BROWN GRIT COVERING BEARING TEMPERATURE SENSOR.
	NON-UNIFORM	CC919	THE BEARING OUTER EDGE DID NOT HAVE VERY UNIFORM APPEARANCE.
	RIDGE	CC894	BEARING DUKE EDGE BROKE; SHOWS RIDGES ON HOUSING
	1.1.1.1.1. Extrusions: blistered, blemished, burred, with whiskers, with nodules, with extrusions, with protrusions, with asperities, with dendritic growth, built up, fuzzy, with flakes		
	BLISTER	CC919	BEARING ANTERIOR SUPPORT FAILED; SHOWING BLISTERED APPEARANCE.
	WITH NODULES	CC509	BROKEN WASHERS/SEALS IN FUEL PUMP SHAFT BEARING; WITH NODULES 7/8"
	1.1.1.1.2. Gaps: pitted, dented, dinged, gouged, cratered, with voids, with cavities, with vesicles, with holes, with gaps, with indentations, with craters, permeable, porous		
	PIT	CE423	WEAR UNDER BEARING ASSY ATTACH PLATE, PITTING ON L EDGE
	WITH CAVITIES	CE423	IRREGULARITIES WITH CAVITIES SHOWN BY INSPECTION OF SPHERICAL BEARING

Fig. 1. *Clustering of eight uneven bearing cases into a class and two subclasses.*

Aerospace risks, problems, and corrective actions involve hardware, software, processes, operations, humans, and organizations. Analysis of discrepancies, problems, and mishaps typically involves considering a wide spectrum of potential causes, from mechanical to organizational domains. Key information about a risk, problem, or discrepancy is commonly embedded in short titles or disposition summaries. Authors embed key information in free text fields because it can be hard to identify the right data codes. The codes can be confusing or out of date, and they may not match the problem situation being reported. Individual keywords can be extracted from the text as tags, but simple keyword approaches are as brittle as codes. There are too many ways of conveying an idea in natural language. Statistical text mining (a learning approach) is sometimes used to overcome this problem by finding clusters of related terms based on text co-occurrences among data records or documents. This approach can be applied when very little is known about the text. A serious drawback is that experts need to work to make sense of the clusters. This can be difficult and frustrating work.

To solve this problem, engineers at Johnson Space Center (JSC) developed an alternative text analysis approach called semantic parsing, which uses taxonomies to tag and categorize key information in text descriptions of aerospace entities, functions, properties, and problems. The taxonomies are enhanced with some ontology-based relationships. Taxonomies define concepts and concept hierarchies. Ontologies include taxonomies and add value by using axioms, restrictions, and properties to relate concepts. They not only accommodate many ways of expressing concepts, but the concepts in ontologies can also be combined (via relations between categorized terms) into phrases that can be extracted from text and analyzed. Text parsing that uses ontologies is thus "semantic." The semantic parsing approach requires development and

maintenance of taxonomies and ontologies. Engineering complex aerospace systems involves many heterogeneous disciplines, and incorporating heterogeneous domains and viewpoints is a significant challenge.

Engineers at JSC have developed a comprehensive aerospace ontology to support several projects in the last few years. Aerospace ontology was first developed for extracting information from text for hazard and risk analysis. The semantic text parsing approach was applied to requirements text and risk data, to support semi-automated generation of system interaction models from text. System structure (functional or physical), vulnerabilities, and safeguards were captured for analysis. Equipment classes and Function/Action classes were also used to organize default information (i.e., standard hazards, vulnerabilities, and safeguards) for system components in libraries. Defaults can guide users to identify potential system problems and mitigations. Recently, the work has focused on identifying recurring problems in problem and discrepancy report databases. The goal is to classify problems (i.e., hazard, damage, impairment, or discrepancy) and the things that have the problems. The semantic parser finds objects or occurrences that have problematic attributes and uses the classification of these things to browse and search reports. Several browsing interfaces have been prototyped. Figure 1 shows an example interface for browsing "uneven bearings" and shows the class "Uneven" and subclasses "Extrusions" and "Gaps" with their associated mapping words (i.e., words that map to the class when found in text).

An advantage of semantic tagging is that it can be combined with other tagging and text-mining approaches and tools. Combining semantic and statistical approaches with faceted browsing should lead to the powerful tool that NASA needs for extracting knowledge and models for text specifications and finding patterns in problem reports. The next step would be to move from operational prototypes to integration into a commercial off-the-shelf-based text analytic implementation for robust use.