The 2016 AAAI Fall Symposium Series:
Cognitive Assistance in Government and Public Sector Applications
Technical Report FS-16-02

Cognitive Cyber Security Assistants – Computationally Deriving Cyber Intelligence and Course of Actions

[1]Charles Palmer, [2]Lee Angelelli, [3]Jeb Linton, [4]Harmeet Singh, [5]Michael Muresan

[1]CTO Security and Privacy, Distinguished Research Staff Member, IBM Research, ccpalmer@us.ibm.com

[2]IT Architect, IBM US Federal CTO Office, LAngelel@us.ibm.com

[3]Watson Chief Security Architect, STSM IBM Watson, jrlinton@us.ibm.com

[4]IT specialist IBM Cloud, harsingh@us.ibm.com

[5]Executive Architect IBM Cloud, muresan@us.ibm.com

Abstract

Cyber security organizations charged with protecting IT infrastructures face daunting complex challenges. These include a vast array of information sources of variable trustworthiness, overwhelming numbers of incident reports, quickly changing offensive and defensive tactics, and the ongoing shortage of skilled cybersecurity personnel. Cognitive systems offer a new approach to addressing these challenges. Using natural language processing and machine learning techniques, systems such as IBM Watson™ can incorporate an enormous amount of information each day, discover the entities and relationships described, and apply reasoning over that knowledge to understand questions from the Security Analyst and provide answers in their own language. In this paper we discuss the technologies and techniques we used to build a cognitive cyber security assistant.

Introduction

Cybersecurity has always been hard. With our growing use of computers and communications in everything from critical systems to national defense, including mobile devices driving an ever-widening enterprise perimeter, this cyber security challenge grows even faster. Hiring, training, and retaining cyber security personnel to support this growth has always been difficult, but it is now becoming critical.

What makes a good cyber security analyst? If you talk to them, you will find fast readers with broad computer and networking knowledge and a great memory. That's because to keep up in cyber security today an analyst has to read – a lot – every day: new advisories, news articles, threat analyses, patches, incident reports, blogs, and information from within their own organization and their industry or sector peers. All of this needs to be read and understood, in addition to performing all their other duties. Finding the time to really digest and incorporate this daily flow of new knowledge with what they had already learned is slow, difficult and can lead to mistakes. As the number and complexity of the indispensable systems we depend on grows, while the number of cyber security analysts grows less quickly, if at all, we have a scalability problem. During times of conflict, this scalability problem quickly becomes evident, to both the defender and the attacker.

The information the analyst needs to digest is unstructured natural language written by humans for humans. Until now, computers were not particularly adept at processing unstructured natural language. Since more than 80% of all the information being produced worldwide each day is unstructured, that data has been called "dark data" since our computers couldn't "see it" (Kelly and Hamm, 2013).

Another challenge is that much of the meaning carried by natural language is implicit – the exact meaning is not completely and precisely stated. It can be highly dependent on the context, such as what was said before, the current topic or environment (e.g., peace or war), or who is discussing it. Moreover, natural language is typically imprecise -- it doesn't treat a subject with repeatable, numerical precision. Humans are always dealing with varying degrees of uncertainty and fuzzy associations between words and concepts. Humans also can develop a pretty good filter for phony or suspect information. Thus, humans bring a huge amount of background knowledge to reconcile these inconsistencies and interpret what they read while making connections to what they have read before.

The leverage that cognitive systems provide is aimed precisely at resolving these issues. These systems apply big data analytics, natural language processing, machine learning, graph computing technologies and more to:

- Understand natural language text using Natural language processing, and to communicate with humans in their own language;
- Ingest the huge volume of unstructured cyber security data generated every day, such as advisories, news articles, threat analyses, patches, incident reports, blogs, and internal/external shared information;
- Extract tacit and explicit knowledge from unstructured text and incorporate it into the knowledge already gained.

This paper will discuss how cognitive technologies can enable systems to understand and correlate adversaries' intents, the tactics-techniques-procedures (TTPs), used to exploit targeted victims' vulnerabilities, and the forensic data associated with a specific cyber attack or campaign.

Cognitive Cyber Security Assistant Capabilities

The cognitive cyber security assistant combines three main technologies: NLP/Information Extraction, hypothesis generation and evaluation, and dynamic learning computing to effectively harness the explosion of unstructured data (Kelly and Hamm 2013).

These cognitive technologies allow the system to ingest and identify entities and relationships from cyber security data sources. The results are stored in a cyber intelligence corpus and which includes a knowledge graph, as the example in Figure 1 shows.

Figure 1. Knowledge graph.

Humans and Cognitive Systems Interactions

Kelly and Hamm observed that "The goal isn't to replicate human brains, though. This isn't about replacing human thinking with machine thinking. Rather, in the era of cognitive systems, humans and machines will collaborate to produce better results, each bringing their own superior skills to the partnership. The machines will be more rational and analytic—and, of course, possess encyclopedic memories and tremendous computational abilities. People will provide expertise, judgment, intuition, empathy, a moral compass, and human creativity. … cognitive systems will be designed to draw inferences from data and pursue the objectives they were given." (Kelly and Hamm 2013)

To enable effective interaction between the cognitive assistant and human analyst, the cognitive cyber security assistant must go beyond a keyword search paradigm. It must use natural language to adapt the human–machine interface.

Use case

A typical use case for a cognitive cyber security system is detecting multiple non-obvious cyber attacks in various stages occurring over some time period in different geographic regions. The system uses its corpora to hypothesize and draw evidence-based conclusions. For example, the combination of specific spear phishing emails (attack delivery), the Upatre trojan establishing command and control (C2), and distributed denial of service (DDOS) attacks (deception) that are executed while the Dyre malware exfiltrates money from victims, indicates that the Dyre Wolf (IBM1 2015) campaign is being launched against an organization(s). Cognitive cyber security solutions are able to extract characteristics of the threat actors, TTPs employed, exploited targets, and intended effect(s). Cognitive cyber security solutions can also be used to combine an understanding of an organization's IT infrastructure and potentially-related known vulnerabilities to identify likely vulnerable systems and possible impacts to guide the analyst to a properly prioritized list of actions.

How Cognitive Systems Understand Natural Language

A cognitive cyber security system can bring cyber security defense in depth awareness and intelligence to a level previously unattainable by classical security systems. Such systems may implement deep semantic reasoning in the form of natural language machine learning technologies, in order to "understand" natural human language. By ingesting cyber security documents such as security lab reports, news feeds, Wikipedia; normalizing, extracting, and representing the ontologies (form) and relationships

(function) of relevant entities; these systems perform a continuous accumulation of cyber security intelligence. For cognitive systems to assist humans in understanding real-world cyber security problems and reasoning to determine the most appropriate action, there are many entities, behaviors and interactions that involve human thinking. The cognitive cyber security assistant described here uses the integrated technologies described below, known as the DeepQA architecture, emulating human thinking to enable security professionals to ask questions in natural human language and receive direct, confidence-based responses.

A wide range of Natural Language Processing (NLP) techniques such as Part of Speech Tagging and Entity/Relationship Extraction are used to annotate the corpus of text heavily as it is gathered and ingested. Many of these annotations are domain-specific, or specific to the many Scorer analytics that will be used to rank candidate answers when the system is queried. The resulting annotated corpus is indexed for natural-language search using SOLR and the index cached in memory in order to speed query responses.

Once the annotated corpus is loaded, the DeepQA Factoid Pipeline can be used to answer questions against it. Figure 2 shows the composition of the Factoid Pipeline.

Figure 2. IBM Watson's DeepQA Factoid Pipeline

When a question is posed, Watson performs the following steps in order to find an answer.

- ***Question Analysis*** – The system analyzes the linguistics of the question by decomposing the question using deep parsing and analyzing the question to understand what is being asked and what constraints are being imposed on the answer. Technologies like named entity recognizers (NERs) and name entity detectors (NEDs) pull out any recognized people, places, etc.; then using a slot grammar parser (XSG), Watson identifies parts of speech for some of terms, references, conferences, pronouns, detect relationships, etc. in the text. Lexical answer types (LATs) are determined in order to help form candidate answer searches and to score candidate answers by type match to the Focus of the query. Along with the full input text, derived inferences, Focus and LATs will be used to build potential queries in the next phase.
- ***Hypothesis Generation*** – Using output from the Question Decomposition, the system now performs several Lucene searches against the annotated corpus index in order to generate possible answer candidates. This results in typically 100-300 candidate answers (also called Hypotheses), typically in the form of one-word or occasionally longer "factoids". The results of these searches will also be used for scoring and filtering passages in the corpus.
- ***Hypothesis Evidence Scoring*** – Next, all hypotheses are scored by a battery of over 50 different analytics individually and in combination. These "Scorers" score each candidate answer on the basis of grammatical and content-based matching criteria, to build a case for and against each. Criteria include geospatial, gender, temporal, taxonomic, passage support, and source reliability information; various methods of paraphrase and idiomatic language matching, and Type Coercion (TyCor) to ensure that the LAT of the candidate answer matches the LAT of the focus of the question. The array of thousands of ratings from all the scorers against all the candidate answers is provided as input to the next phase.
- ***Final Merge & Ranking*** - At this point, the DeepQA system typically has overlapping or duplicate answers. The system uses feature vectors and normalizing scores derived from the scorers in the preceding step to merge these redundant answers. Once the merge is completed, Watson computes the final confidence scores for the candidate answers by applying a series of machine learning models (primarily using logistic regression) that weight all of the normalized feature vectors scores to produce the final confidence scores.
- ***Supporting Evidence Merging & Ranking*** - This concluding phase retrieves all the evidence that was collected during Primary Search execution; this takes the form of passages of text containing each answer from the original corpus. It applies the Justifying Passage Model to evidence to create a ranked list of Answers & Evidence. Like the primary search – the supporting evidence search uses passage term match, skip bigram, text alignment, and logical form answer candidate scoring.

This ensemble of technologies supports human decision making by mimicking human question answering, in a way that scales to more knowledge and better recall than a human can accomplish. The annotation of the corpus mimics how the human mind evaluates the grammar of a statement and makes connections from it to prior knowledge in memory. Similarly, Question Analysis analyzes the grammar and content of the question, and Hypothesis Generation and Evidence Scoring make connections from the question to the knowledge stored in the corpus – much as the human brain makes and evaluates the strength of connections when evaluating a question. Final Merge and Ranking uses

learned experience to judge the relative value of various means of reasoning to find the best answer, much as a human can use metacognition to evaluate the relative merit of various lines of reasoning. And finally, Supporting Evidence is brought to bear to support the human's evaluation of the cognitive assistant's reasoning, enabling cooperative cognition between human and machine.

By curating the corpus and annotating it with customized annotation algorithms, the system is made capable of understanding and assisting humans in the particular domain of cyber security: helping security professionals to drive new discovery and insight, to better understand the relationships between malicious intent of an attack, types of attacks involved, actors orchestrating the attack, and the actors' methods, tradecraft, and objectives. This process is known as Domain Adaptation.

Domain Adaptation – Understanding the Language of Cyber Security

Domain adaptation is the process of teaching cognitive systems like Watson to understand the entities and relations that are used in a specific domain. The system developers and domain experts use the domain adaptation process to prepare the system to answer questions or provide information. In the current case, the domain is cyber security. Adaptation is an iterative process of experimentation, analysis, and development. The goal of the process is to tailor the system so that it can properly process the corpus of knowledge to provide relevant and meaningful information to the users. In our case, the domain adaption process consisted of these steps:

1. **Corpus creation and curation** – Develop a cyber security corpus consisting of a broad set of information sources which might inform a human expert on the domain. This corpus is likely to contain thousands or millions of articles, blog entries, threat and vulnerability reports, security analyst notes, incident reports, etc. In addition, due to the volatile nature of cyber security, the corpus must be updated daily, if not more often. The corpus of a cognitive cyber security assistant differs from that of a more traditional security tool such as an advanced Security Incident and Event Management (SIEM) tool. The SIEM tool typically uses machine learning techniques to look for anomalous and non-obvious patterns of activity in highly-structured machine-generated information such as system logs and netflow data. It then identifies potentially significant events that a human analyst should investigate. The corpus for the cognitive cyber security assistant consists of ever-changing collections of unstructured text. A subtle difference is that the cognitive assistant is prepared to answer questions that have never been asked before, perhaps discovering previously unknown correlations. This is different from the SIEM tools which are watching for events known *a priori* to indicate a problem. The cognitive security assistant can work in tandem with SEIM tools, complementing them by enabling the human analyst to explore the vast corpus for correlations with what the SIEM system detected.
2. **Pre-Annotation – dictionaries and concept detector** - Dictionaries are used to pre-annotate documents. Dictionaries contain terms related to a domain of interest. The pre-annotator finds terms that are represented in your dictionary and automatically adds annotations for them in the documents. This initial pass on the documents helps the human annotator who has only to accept or correct the pre-annotations while identifying new ones. Dictionaries are assigned an entity based on the domain's ontology (type system) and use lemma, surface forms, and parts of speech to group together words and phrases that identify the same entity. For example, a dictionary name "MALWARE" contains the entry Dyre Wolf (lemma) and Dyreza, dyreza malware, dyreza trojan, etc. A cognitive system like Watson will annotate any of these mentions in an article as MALWARE. In addition, equating these words also benefits information extraction in the surrounding text. For example, what the machine annotator learns from training examples of the texts near "Obama" and "Barack Obama" is applied to texts that the machine annotator sees near other mentions of the US President, because the dictionary states that these terms are equivalent for information-extraction purposes.

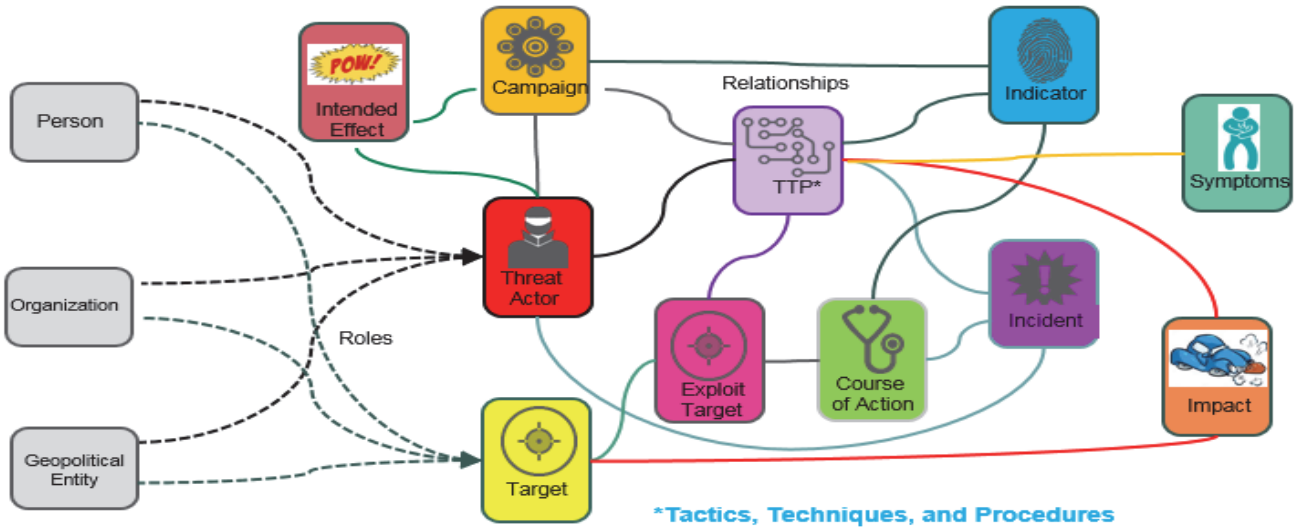

Figure 3. Cyber Watson Type System (Simplified View)

3. **Custom annotator** - The SIRE toolkit (Ahmed 2014) is used to build trainable extractors for new applications and domains. SIRE provides components for entity mention detection using (a) Maximum Entropy models that can be trained from annotated data created by using the Watson Knowledge Studio (Ahmed 2014), (b) a trainable co-reference component for grouping detected mentions in a document that correspond to the same entity, and (c) a trainable relation extraction system. The WKS toolkit produces annotators that can be deployed to a cognitive system ingestion pipeline. Subject matter experts (SMEs) create a type system (ontology) as shown in simplified form in Figure 3; the type system in our case is partly based on the OASIS Structured Threat Information Expression (STIX) standard. The SMEs then manually annotate domain-specific entities, relationships, and coreferences of interest within a document. Once the SMEs have completed annotating a document, the documents are submitted as a ground truth to train and test SIRE models to learn the cyber security domain. The mention detection problem is usually formulated as a classification problem. The Maximum Entropy classification framework integrates arbitrary types of information and makes a classification decision by aggregating all information available for a given classification (Ahmed 2014). The section "Relation Extraction – Extracting Tacit …" will explain SIRE's technical framework.
4. **Ingestion/Training/Answering User Questions** - An important early and recurring step in any Domain Adaptation involves the acquisition and ingestion of content (the corpus) and training of Watson's scoring algorithms and machine learning model against ground truth in the form of representative question/answer pairs. During this phase, Watson learns to classify text passages using logistic regression. Ground truth is divided into training, testing, and blind testing sets, and applied iteratively with statistical analysis of accuracy, precision, confidence, and recall. The trained scorers and machine learning model are subsequently used at run time during question analysis, hypothesis generation, and answer scoring and merging. Answers are evaluated based on the entire state of the corpus and the scoring algorithms at question submission time, and will be consistent until Watson acquires new knowledge, which can be in the form of new (additional or changed) content, iterative subsequent training on new question/answer pairs, or user feedback on the accuracy and usefulness of answers Watson has provided.

Relations Extraction - Extracting Tacit Knowledge from Unstructured Text

Tacit Knowledge (TK) generally refers to information that is difficult to convey, store, or transfer explicitly (Polanyi 1966). In the cyber security domain, the amount of unstructured text information to keep up with new and/or mutating cyber attacks is increasing exponentially. In order to keep up, cyber security organizations must automate the process of extracting cyber security tacit information from unstructured text. For cognitive systems, information extraction is a key technology for developing annotators that

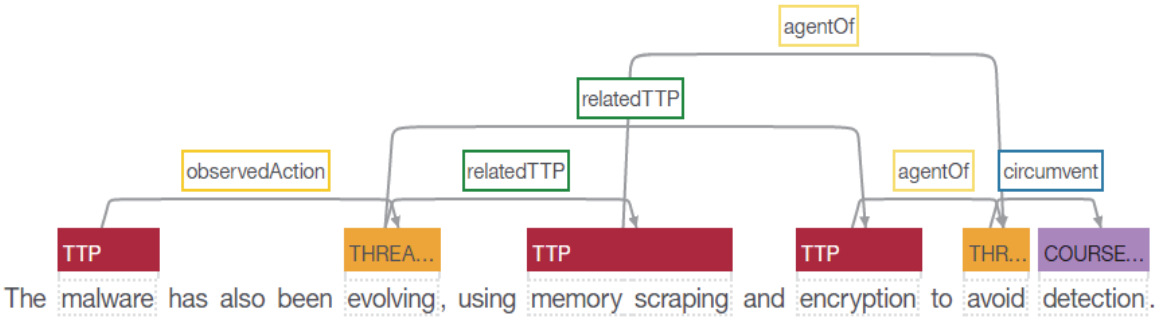

Figure 4. WKS annotation of malware's Related TTPs to avoid detection

understand text, by identifying the important conceptual objects and relations between them in a passage. The technology is meant to build classification and/or detection models. The distinction here is mostly made at decoding time: for "sequence classification", the classification of an example depends on the classification of the surrounding tokens (i.e. POS tagging, text chunking, Named Entity recognition, mention detection), while for "example classification" the examples are not dependent on their surrounding examples (e.g. prepositional phrase attachment, medical diagnosis of a patient, etc). This difference is reflected at decoding time, where search over the possible sequences is needed (e.g., Viterbi or Forward-Backward search) and is not needed for example classification (Radu 2013). The cognitive cyber security assistant used IBM's Statistical Information and Relation Extraction (SIRE), an information extraction technology, which can perform:

- Mention detection: identify spans that are mentions of targeted ontology's entity types (THREAT_ACTOR, VICTIM_TARGETED, TTP, etc.).
- Co-reference resolution: group the mentions within a document that correspond to the same entity.
- Relation extraction: identify relations between pairs of extracted mentions within the same sentence.

Figure 4 illustrates how SIRE leverages the Maximum Entropy classifier (statistical machine translation) to understand cyber security-related entities mentioned, relations between different entities in a sentence, coreferences of the same mentions of the same entities within and across documents, and converts textual data into structured data. Figure 4 shows the WKS machine learning decoded values – where the machine was able to understand a malware [TTP] is evolving [THREAT_RELATED_ACTION] using "memory scraping" [TTP] and "encryption" [TTP] to avoid [THREAT_RELATED_ATION] "detection" [COURSE_OF_ACTION].

Conclusion: Capturing Human Expertise as Labeled Training Data

We've discussed the various ways in which Watson is being trained to become a cognitive cyber security assistant. One can weave these threads together to illustrate more generally how a cognitive assistant should be trained using human expertise as Labeled Data for Supervised Learning. A human Cyber Security SME needs to:

- Recognize cyber security terminology
- Know how cyber security entities relate to each other
- Be able to read, parse, and understand cyber security documents written by others
- Know what questions to ask and where to go for the answers

Notice that the several forms of training we have described in this paper are methods of capturing these forms of human understanding from the experts. Terminology is captured in the form of Dictionaries. Understanding of Entities and Relations is captured in the form of a Type System which is used by humans reading, parsing, and understanding documents via Human Annotation. Thousands of Question-Answer pairs collected from the cyber security SME's are used as the Ground Truth used to train the system to answer questions when assisting the next rising generation of Security Analysts.

Last but not least, the iterative collection and curation of corpus content using feedback from users brings all of these steps together to refine the system's understanding of which content sources are trustworthy. We expect this system to improve greatly with time and iterative feedback through this process. Thus, the system becomes a Cyber Security expert advisor in order to assist humans in becoming experts in turn.

References

Kelly III, John E. 2015. Computing, cognition, and the future of knowing. Webpage retrieved Sept. 7, 2016: http://www.research.ibm.com/software/IBMResearch/multimedia/Computing_Cognition_WhitePaper.pdf.

Kelly III, John E.; Hamm, Steve. 2013. Smart Machines: IBM's Watson and the Era of Cognitive Computing (Columbia Business School Publishing). Columbia University Press. Kindle Edition.

Polanyi, M. 1966. The Tacit Dimension, London: Routledge & Kegan Paul

Ahmed, Dr. Mohamed N. 2014. "Statistical Information and Relation Extraction (SIRE) Toolkit".

IBM1, "The Dyre Wolf". Webpage retrieved Sept. 7, 2016. https://securityintelligence.com/media/research-report-inside-the-dyre-wolf-malware-campaign/ .

Radu Florian. 2013. "How-to Build a MaxEnt Sequence Classification Model with the SIRE Toolkit," IBM TJ Watson Research Center, Yorktown Heights NY.