Machine Learning Extracts Attack Data From Verbose Threat Reports

Originally published at https://www.unite.ai/machine-learning-extracts-attack-data-from-verbose-threat-reports/

New research out of the University of Chicago illustrates the conflict that has arisen in the past ten years between the SEO benefits of long-form content, and the difficulty that machine learning systems have in gleaning essential data from it.

In developing an NLP analysis system to extract essential threat information from Cyber Threat Intelligence (CTI) reports, the Chicago researchers faced three problems: the reports are usually very long, with only a small section dedicated to the actual attack behavior; the style is dense and grammatically complex, with extensive domain-specific information that presumes prior knowledge on the part of the reader; and the material requires cross-domain relationship knowledge, which must be 'memorized' to understand it in context (a persistent problem, the researchers note).

Long-Winded Threat Reports

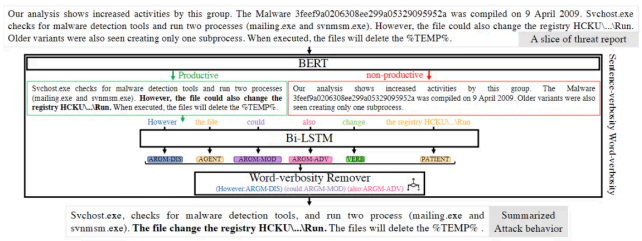
The primary problem is verbosity. For example, the Chicago paper notes that among ClearSky's 42-page 2019 threat report for the DustySky (aka NeD Worm) malware, a mere 11 sentences actually deal with and outline the attack behavior.

The second obstacle is text complexity, and, effectively, sentence length: the researchers observe that among 4020 threat reports from Microsoft's threat report center, the average sentence comprises 52 words – only nine short of the average sentence length 500 years ago (in the context of the fact that sentence length has declined 75% since then).

However, the paper contends that these long sentences are essentially 'compressed paragraphs' in themselves, full of clauses, adverbs and adjectives that shroud the core meaning of the information; and that the sentences often lack the basic conventional punctuation which NLP systems such as spaCy, Stanford and NLTK rely on to infer intent or extract hard data.

NLP To Extract Salient Threat Information

The machine learning pipeline that the Chicago researchers have developed to address this is called EXTRACTOR, and uses NLP techniques to generate graphs which distill and summarize attack behavior from long-form, discursive reports. The process discards the historical, narrative and even geographical ornamentation that creates an engaging and exhaustive 'story' at the expense of clearly prioritizing the informational payload.



Source: https://arxiv.org/pdf/2104.08618.pdf

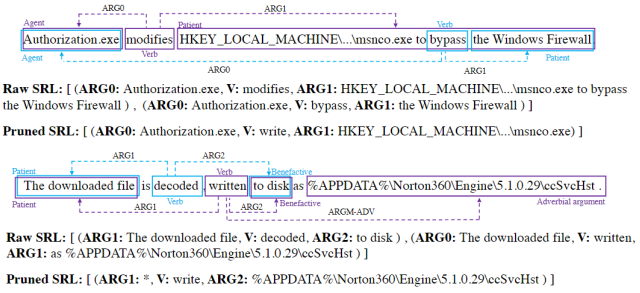
Since context is such a challenge in verbose and prolix CTI reports, the researchers chose the BERT (Bidirectional Encoder Representations from Transformer) language representation model over Google's Word2Vec or Stanford's GloVe (Global Vectors for Word Representation).

BERT evaluates words from their surrounding context, and also develops embeddings for subwords (i.e. *launch*, *launching* and *launches* all stem down to *launch*). This helps EXTRACTOR to cope with technical vocabulary that is not present in BERT's training model, and to classify sentences as 'productive' (containing pertinent information) or 'non-productive'.

Increasing Local Vocabulary

Inevitably some specific domain insight must be integrated into an NLP pipeline dealing with material of this kind, since highly pertinent word forms such as IP addresses and technical process names must not be cast aside.

Later parts of the process use a BiLSTM (Bidirectional LSTM) network to tackle word verbosity, deriving semantic roles for sentence parts, before removing unproductive words. BiLSTM is well-suited for this, since it can correlate the long-distance dependencies that appear in verbose documents, where greater attention and retention is necessary to deduce context.



EXTRACTOR defines semantic roles and relationships between words, with roles generated by Proposition Bank (PropBank) annotations.

In tests, EXTRACTOR (partially funded by DARPA) was found capable of matching human data extraction from DARPA reports. The system was also run against a high volume of unstructured reports from Microsoft Security Intelligence and the TrendMicro Threat Encyclopedia, successfully extracting salient information in a majority of cases.

The researchers concede that the performance of EXTRACTOR is likely to diminish when attempting to distill actions that occur across a number of sentences or paragraphs, though re-tooling the system to accommodate other reports is indicated as a way forward here. However, this is essentially falling back to human-led labeling by proxy.

Length == Authority?

It's interesting to note the ongoing tension between the way that Google's arcane SEO algorithms seem to have increasingly rewarded long-form content in recent years (although official advice on this score is contradictory), and the challenges that AI researchers (including many major Google research initiatives) face in decoding intent and actual data from these increasingly discursive and lengthy articles.

It's arguable that in rewarding longer content, Google is presuming a consistent quality that it isn't necessarily able to identify or quantify yet through NLP processes, except by counting the number of authority sites that link to it (a 'meatware' metric, in most cases); and that it is therefore not unusual to see posts of 2,500 words or more attaining SERPS prominence regardless of narrative 'bloat', so long as the extra content is broadly intelligible and does not breach other guidelines.

Where's The Recipe?

Consequently, word counts are rising, partly because of a genuine desire for good long-form content, but also because 'storifying' a few scant facts can raise a piece's length to ideal SEO standards, and allow slight content to compete equally with higher-effort output.

One example of this is recipe sites, frequently complained of in the Hacker News community for prefacing the core information (the recipe) with scads of autobiographical or whimsical content designed to create a story-driven 'recipe experience', and to push what would otherwise be a very low word-count up into the SEO-friendly 2,500+ word region.

A number of purely procedural solutions have emerged to extract actual recipes from verbose recipe sites, including open source recipe scrapers, and recipe extractors for Firefox and Chrome. Machine learning is also concerned with this, with various approaches from Japan, the US and Portugal, as well as research from Stanford, among others.

In terms of the threat intelligence reports addressed by the Chicago researchers, the general practice of verbose threat reporting may be due in part to the need to reflect the scale of an achievement (which can otherwise often be summarized in a paragraph) by creating a very long narrative around it, and using word-length as a proxy for the scale of effort involved, regardless of applicability.

Secondly, in a climate where the originating source of a story is often lost to bad citation practices by popular news outlets, producing a higher volume of words than any re-reporting journalist could replicate guarantees a SERPS win by sheer word-volume, assuming that verbosity – now a growing challenge to NLP – is really rewarded in this way.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More