CyberEnt: Extracting Domain Specific Entities from Cybersecurity Text

Casey Hanks, Michael Maiden, Priyanka Ranade, Tim Finin, and Anupam Joshi
University of Maryland, Baltimore County
Baltimore, MD 21250
{chanks1, mmaiden1, priyankaranade, finin, joshi}@umbc.edu

Entity Recognition is a critical component of automated knowledge extraction, allowing language understanding models to label instances of real-world entities in text. To accomplish this, Natural Language Processing (NLP) models must be trained on very large corpora of human-annotated text. There are many domain-agnostic text corpora available for training models on generic entity types such as *Person, Organization*, and *Date*. However, general domain entity types are not sufficient for more specialized fields like *cybersecurity* because they are unable to recognize cybersecurity-specific entities such as malware-type, operating system, or attack-type, which can be useful in downstream tasks such as malware analysis, attack and vulnerability classification, and Cybersecurity Knowledge Graph (CKG) completion. (Mulwad et al., 2011; Joshi et al., 2013; Gao et al., 2021; Georgescu et al., 2021).

There is an ever-growing volume of cyber threat intelligence (CTI) available online, making it increasingly difficult for human analysts to sift through and use. As a result, there is a large need to develop community-accessible datasets to train existing AI-based cybersecurity pipelines to efficiently and accurately extract meaningful insights from CTI. Unlike fields like medicine or law, cybersecurity has few comprehensive training datasets that are available *and* continuously updated.

We have created an initial large, unstructured CTI corpus from a variety of open sources such as cybersecurity vendor reports/blogs, vulnerability databases (Common Vulnerabilities and Exposures (CVE)) records, and Advanced Persistent Threat (APT) reports. We are using the corpus to train and test cybersecurity entity models using the SpaCy (Honnibal et al., 2020) framework and in particular, exploring self-learning methods to automatically recognize cybersecurity entities, based on limited, but high-quality training datasets.

Malware_Name	Campaign
Malware_Type	IP_Address
Software_Name	Protocol
Version_Tag	Threat_Actor
Vulnerability	Operating_System
Attack_Type	Hash
Programming_Language	URL

Table 1: Additional domain specific entity types

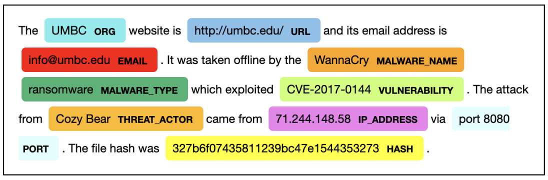
The UMBC ORG website is http://umbc.edu/ URL and its email address is info@umbc.edu EMAIL . It was taken offline by the WannaCry MALWARE_NAME ransomware MALWARE_TYPE which exploited CVE-2017-0144 VULNERABILITY . The attack from Cozy Bear THREAT_ACTOR came from 71.244.148.58 IP_ADDRESS via port 8080 PORT . The file hash was 327b6f07435811239bc47e1544353273 HASH .

Figure 1: Cybersecurity text with entities found

Table 1 shows the current list of cybersecurity entity types we support in addition to SpaCy's OntoNotes types. We take a *data-driven* approach when creating our training dataset by using intensive evaluation criteria for entity recognition annotations. Our initial annotated dataset was reasonable but ultimately resulted in an unsatisfactory evaluation for training a cybersecurity-based entity recognition model. In addition, we are exploring tools available in the SpaCy NLP framework. Examples include regex-based recognition (for entities like URLS, IP addresses, hash values, and CVE identifiers), and SpaCy's *entityRuler* tool used to recognize and train on names of instances of types like operating systems extracted from Wikidata.

In addition to refining the training dataset, our future work will survey and test SpaCy NLP tools, and create methods for continuous integration of new information. In particular, we will add a *coreference module* and link typed entities to Wikidata items. The cybersecurity entities, relations, and events will be used to populate and extend existing CKGs (Satyapanich et al., 2020; Piplai et al., 2020; Mitra et al., 2022). Lastly, we will employ continuous, periodic web-scraping and pulling from open-source CTI feeds to integrate new data.

Acknowledgement. This material is based upon work supported by a grant from NSA and from National Science Foundation Grant No. 2114892.

References

Chen Gao, Xuan Zhang, Mengting Han, and Hui Liu. 2021. A review on cyber security named entity recognition. *Frontiers of Information Technology & Electronic Engineering*, 22(9):1153–1168.

Tiberiu-Marian Georgescu, Bogdan Iancu, Alin Zamfiroiu, Mihai Doinea, Catalin Emilian Boja, and Cosmin Cartas. 2021. A survey on named entity recognition solutions applied for cybersecurity-related text processing. In *Proceedings of Fifth International Congress on Information and Communication Technology*, pages 316–325, Singapore. Springer Singapore.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python. https://spacy.io/.

Arnav Joshi, Ravendar Lal, Tim Finin, and Anupam Joshi. 2013. Extracting cybersecurity related linked data from text. In *Seventh International Conference on Semantic Computing, Irvine, CA, USA, September 16-18, 2013*, pages 252–259. IEEE Computer Society.

Shaswata Mitra, Aritran Piplai, Sudip Mittal, and Anupam Joshi. 2022. Combating Fake Cyber Threat Intelligence using Provenance in Cybersecurity Knowledge Graphs.

Varish Mulwad, Wenjia Li, Anupam Joshi, Tim Finin, and Krishnamurthy Viswanathan. 2011. Extracting information about security vulnerabilities from web text. In *Proceedings of the 2011 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops*, pages 257–260. IEEE Computer Society.

Aritran Piplai, Sudip Mittal, Anupam Joshi, Tim Finin, James Holt, and Richard Zak. 2020. Creating Cybersecurity Knowledge Graphs from Malware After Action Reports. In *IEEE Access Journal*, volume 8.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. CASIE: extracting cybersecurity event information from text. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8749–8757. AAAI Press.