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.

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


