CyberEnt: Extracting Domain Specific Entities from Cybersecurity Text

Casey Hanks, Michael Maiden, Priyanka Ranade, Tim Finin, Anupam Joshi
University of Maryland, Baltimore County

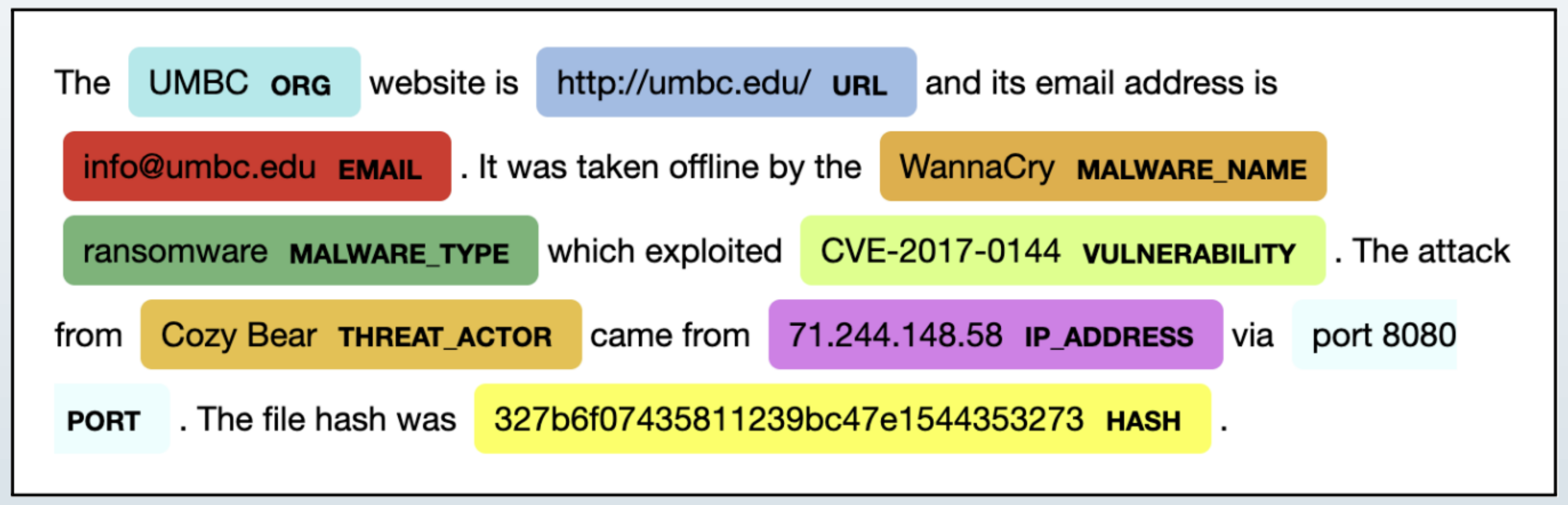
Introduction

What is NLP?

- Natural Language Processing (NLP) is the way in which a computer understands human language.
- Entity Recognition is how a computer can identify and categorize certain words.

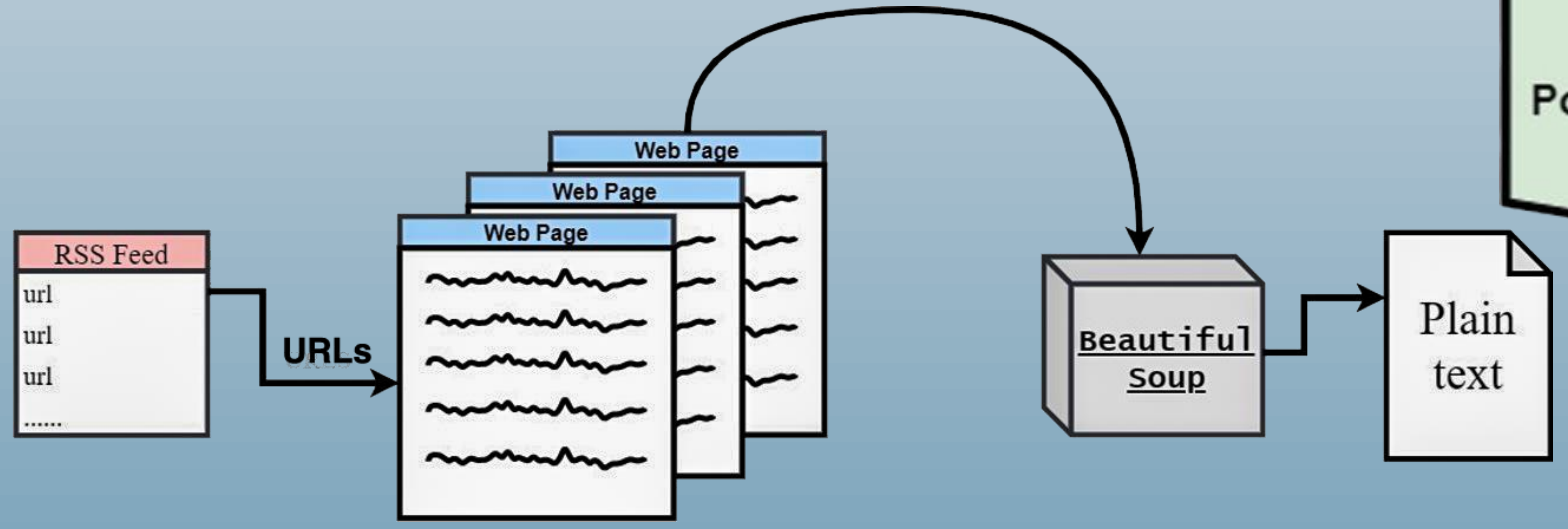
How will we use it?

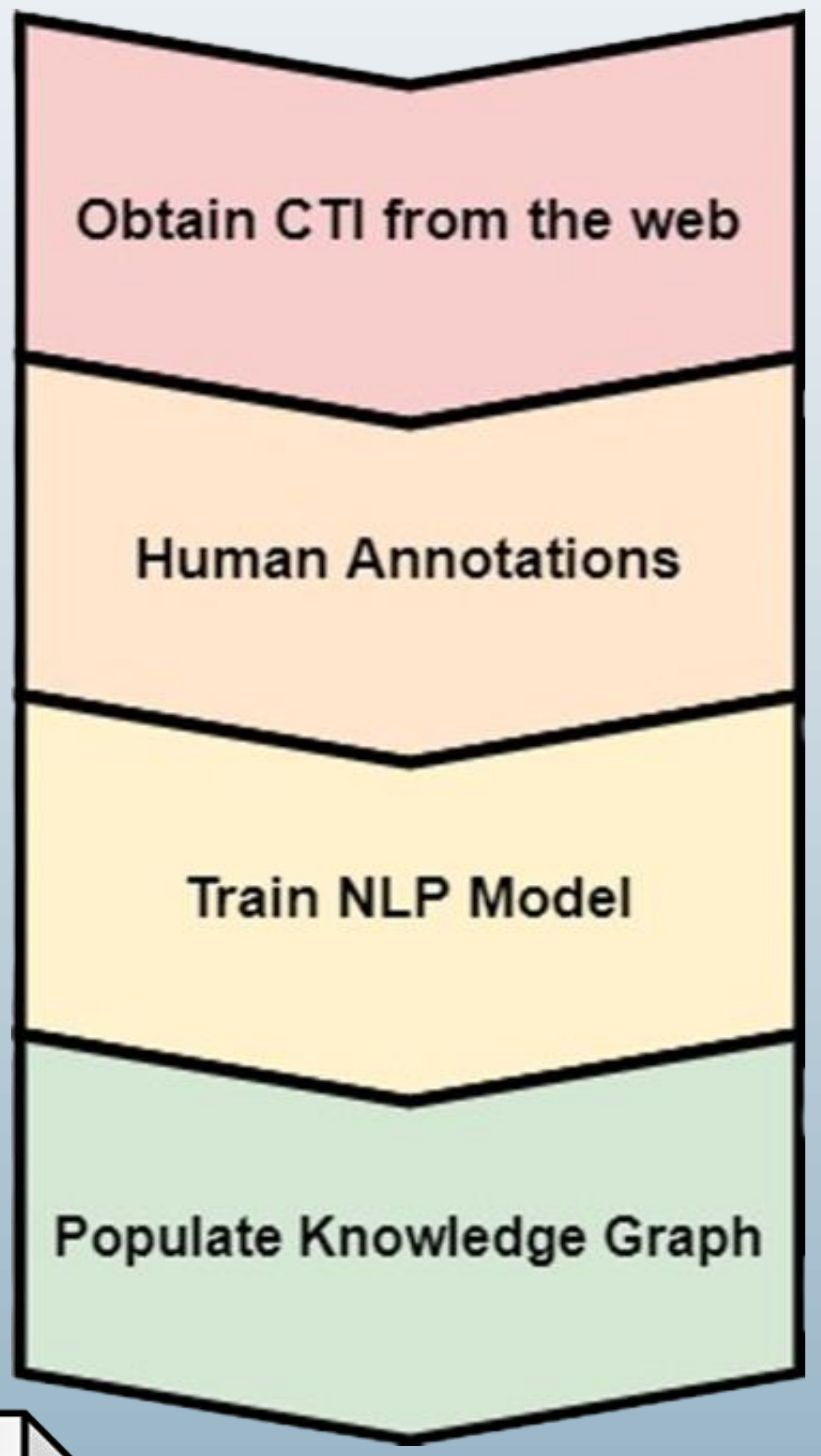
- Training a computer model to categorize certain words pertaining to the cybersecurity field using NLP.
- The model is trained with a large amount of human labeled data.
- There are a lot of labeled data sets for general use but there is a very limited amount for cybersecurity use.
- Teaching the computer to recognize cybersecurity entities is useful for many different purposes like malware analysis.



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 .

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





Methodology

Tasks that we have accomplished

- Building a collection of cybersecurity text that obtains the newest articles from a variety of sources using code
- Updating this collection regularly
- Determining the cybersecurity related categories that the computer will to be able to recognize.
- Training annotators to create human labeled data.
- The annotation of over 1000 sentences for training an NLP model
- Training an NLP model to determine the quality of our annotations

Preliminary Results

- The model had an accuracy of about 65% which is fair but ultimately unsatisfactory.
- We believe this to be due to multiple factors
  - Annotators labeling the same word with different labels or different precision.
  - Several entity types had very low volumes of annotation
  - Lower quantity of annotations than expected
- After annotating over 1300 sentences, only about 400 of them contained annotations

Next Steps

Finding ways to reduce error and improve accuracy

- Revise our list of categories so that annotators have less trouble classifying terms.
- A more in-depth training session with more explanations and live examples
- Taking advantage of other tools that could be used to aid the annotation process.
  - The SpaCy Entity Ruler tool which allows users to make a list of words under each category and automatically have these words labeled.
  - It also allows users to use rule-based methods to automatically categorize certain terms within the text that follow a certain pattern, for example emails and IP addresses

Current work

- With these improvements in place, we are doing another round of annotation, this time with over 2000 sentences

Future Work

- Develop methods for the continuous integration of new information.
- Information from the model will be used to populate a cybersecurity knowledge base.

Acknowledgements

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