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.

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

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

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

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.