Bootstrapping for Text Extraction in Cyber Security

Nikki McNeil[1], Robert Bridges[2], and John Goodall[3]
[1]University of Maryland, Baltimore County
[2,3] Computational Sciences and Engineering Division, Oak Ridge National Laboratory
{nmcneil1@umbc.edu, rabridges@ornl.gov, jgoodall@ornl.gov}

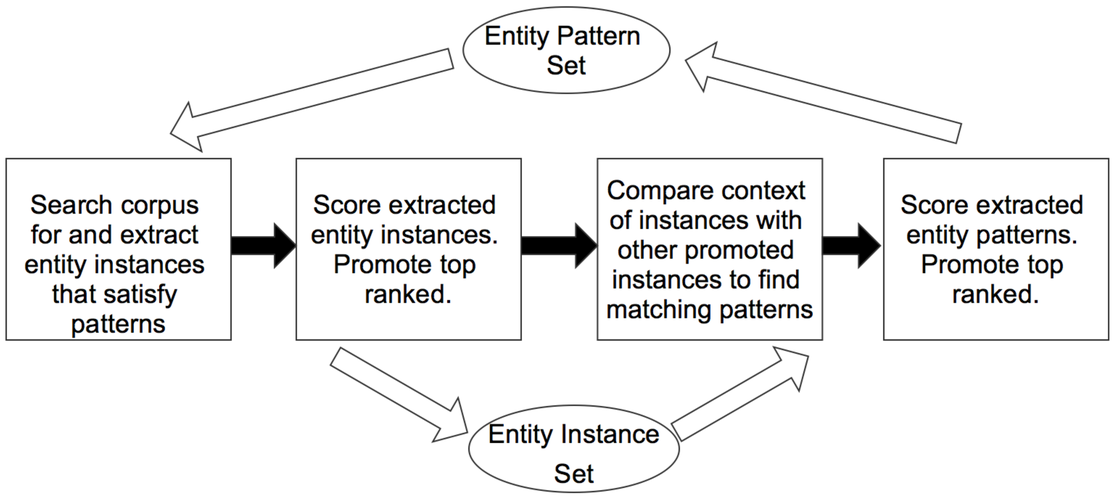
OAK RIDGE National Laboratory

Objective

To extract information about cyber security threats from relevant unstructured text sources - such as news sites, mailing lists, and social media - using the process of bootstrapping.

Bootstrapping Method

Given an ontology and seed instances and patterns:



Advantages of Bootstrapping

- Requires only a small set of labeled data
- Identical process for entity and relation extraction

Example Output

Instance Name	Type	Prefix	Suffix	Document	Score
execute arbitrary programs	Vulnerability Potential Effects	Internet Explorer allows attackers to	" . Heise Media UK.	http://en.wikipedia.org/wiki/Cross-site_scripting	5.7004
Adobe Flash Player	Vulnerability Software	HTTP header injection vulnerabilities in	" . Adobe Systems. November	http://en.wikipedia.org/wiki/Cross-site_scripting	5.7279
cross-site request forgery	Vulnerability Category	) is the equivalent of	(CSRF) in desktop	http://en.wikipedia.org/wiki/Cross-application_scripting	5.6724

Improvements

- Two types of patterns: a phrase to itself be extracted as an instance, or a prefix/suffix that indicates a nearby phrase to be extracted
- Stem all words in patterns
- Phrase is extracted if it contains, rather than matches, all tokens of a pattern
- Instead of the nouns extracted in other domains, extract longer phrases ending in nouns

Basilisk Scoring Method

Score for Patterns:

$$\frac{F_i}{N_i} * \log_2(F_i)$$

The number of promoted instances that occur with the pattern divided by the total number of times the pattern is seen in the corpus

Score for Instances:

$$\sum_{j=1}^{P} \frac{\log_2(F_j + 1)}{P}$$

The total number of promoted instances that have been extracted by all patterns which this instance satisfied, divided by the number of patterns

Conclusion

Scoring Method	Corpus	Number of Iterations	Results
Based on pattern specificity	240 Wikipedia articles	3	32% precision
Basilisk	240 Wikipedia articles	3	42% precision
Basilisk	10 news articles	indefinite	64% precision, 38% recall, .48 F-score

Results are dependent on scoring method, types of documents, and seeds used.

Future goals are to utilize more useful seeds and relation extraction.

The authors would like to thank Professor Bogdan Czejdo and Nicolas Perez of Fayetteville State University and Michael Iannacone of Oak Ridge National Laboratory for the domain exptertise they contributed to this project. This research was performed under an appointment to the U.S. Department of Homeland Security (DHS) Science & Technology (S&T) Directorate Office of University Programs HS-STEM Summer Internship Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and DHS. ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-AC05-06OR23100. All opinions expressed in this poster are the author's and do not necessarily reflect the policies and views of DHS, DOE, or ORAU/ORISE.