Text Analysis for Timely Discovery of Cyber Security Concepts

Corinne Jones[1], Robert Bridges[2], Mike Iannacone[2], and John Goodall[2]
[1] The Pennsylvania State University
[2] Oak Ridge National Laboratory
clj180@psu.edu, {bridgesra, iannaconemd, jgoodall} @ ornl.gov

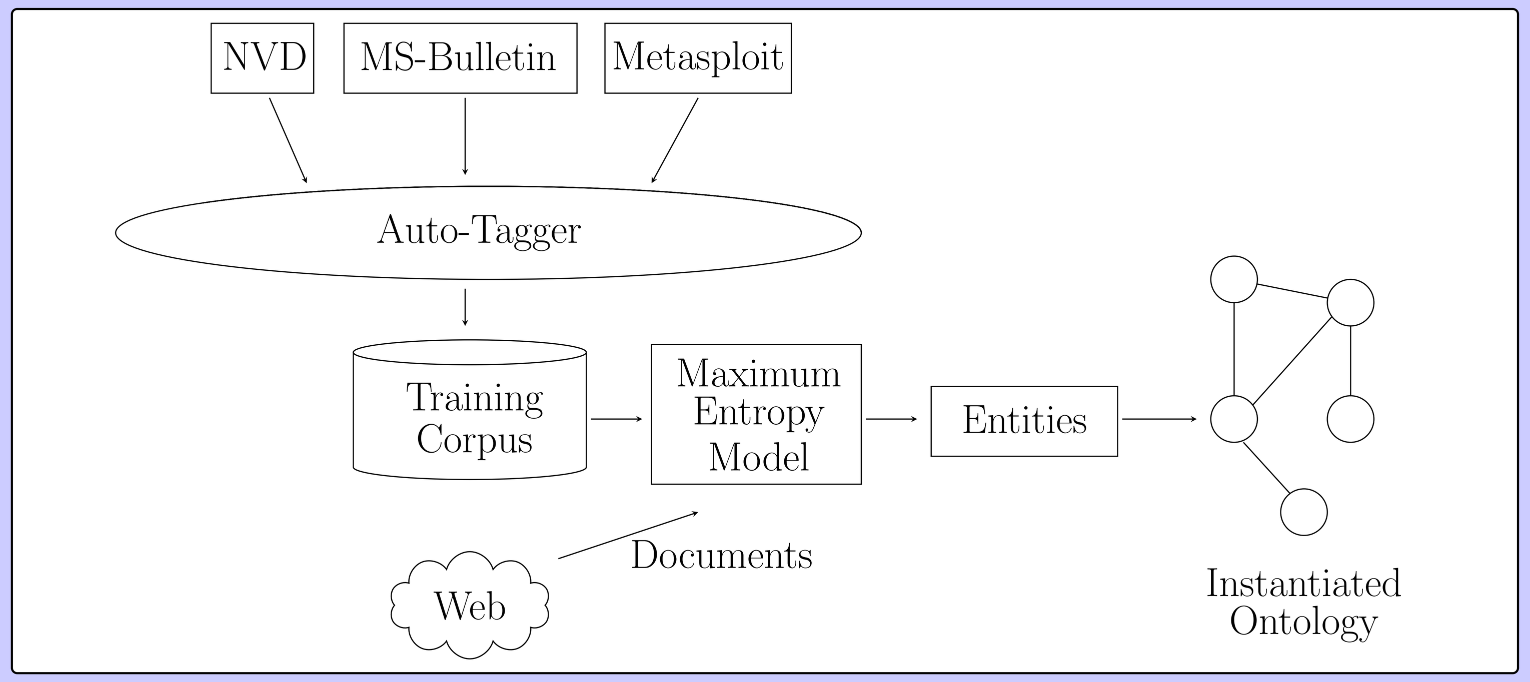
Situation Awareness Visual Analytics

Overview

- Goal: Timely discovery of cyber security concepts
- Method: Entity extraction
- Problem: No labeled data exists in this domain

Approach

1. Gather data
2. Auto-tag (unsupervised learning)
3. Implement maximum entropy model (supervised learning)



Entity Extraction Results

Results from using OWL-QN with l^1 regularization and the Collins Perceptron:

	OWL-QN	Perceptron	Perceptron
Precision	90.5%	99.0%	94.2%
Recall	93.6%	77.3%	96.4%
F-score	92.0%	86.8%	95.3%
Accuracy	94.5%	91.8%	96.8%
λ	1	N/A	N/A
n	2,500	2,500	15,192

Here n is the number of training and test sentences (split 80%/20%)

Auto-Tagging (Unsupervised Learning)

We use structured data along with regular expressions and a relevant terms list to tag unstructured text.

EXAMPLE (NVD):

- ID: CVE-2013-1012
- CPE VECTOR: cpe:/a:apple: safari:6.0.4 and previous versions
- TEXT: Cross-site scripting (XSS) vulnerability in WebKit in Apple Safari before 6.0.5 allows remote attackers to inject arbitrary web script or HTML via vectors involving IFRAME elements.

ENTITIES IDENTIFIED:

- Platform Entities:
  - Vendor
  - Software
  - Hardware
  - OS
  - Version
- Vulnerability Entities:
  - CVE-ID
  - Files
  - Functions
  - Relevant Terms

Auto-Tagging Results

Results acquired via manual inspection of 25 randomly selected descriptions from each source (Over **850,000** words tagged!):

- NVD:
  - Precision $\equiv \frac{\text{\# words correctly labeled}}{\text{Total \# words labeled}} = 100\%$
  - Recall $\equiv \frac{\text{\# words correctly labeled}}{\text{\# words that should be labeled}} = 81\%$
  - F-score $\equiv \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} = 89.5\%$
- MS-Bulletin:
  - Precision = 99.4%
  - Recall = 75.3%
  - F-score = 85.7%
- Metasploit:
  - Precision = 95.3%
  - Recall = 54.3%
  - F-score = 69.1%

Acknowledgements

The DHS Science & Technology, Cyber Security Division provided funding for this project.

History-Based Model

Advantages:

- Robust feature selection
  - E.g., $f_1(x,y) = \begin{cases} 1 & \text{if } t_i = \text{B:vendor}, w_{i-1} = \text{"the"} \\ 0 & \text{else} \end{cases}$
- More sophisticated probability model
- Maximizes entropy

Model:

$$p(t_i|t_{i-2}, t_{i-1}, w_{i-2}, w_{i-1}, w_i) \equiv \frac{e^{f(\bar{t}_i, \bar{w}_i)\cdot v}}{z(\bar{t}_i, \bar{w}_i)},$$

where $z(\bar{t}_i, \bar{w}_i) \equiv \sum_{\hat{t}} \exp[f(t_{i-2}, t_{i-1}, \hat{t}, \bar{w}_i)\cdot v]$.

Goal: Maximize the regularized log-likelihood of the conditional probability of the training examples:

$$L(v) = \sum f(w,t)\cdot v - \sum \log Z(w) - \lambda \sum |v_i|$$

Concave and unique maximum $\implies$ Can use a quasi-Newton method such as OWL-QN

Heuristic Approach

Advantages of the Perceptron:

- Less computationally intensive
- Does not make assumptions of history-based model

Goal: Find the best values for the parameter vector by looping over each training set example several times and updating v at each step k:

$$v_k = v_{k-1} + f(w,t) - f(w,t'),$$

where t is the "gold standard" tag sequence for w given by the training data and t' the most probable tag sequence for w given by v_{n-1}.

Entity Extraction

Goal: Find the optimal tag sequence, $t^*_{[1:n]}$, for each input sentence by solving

$$\arg\max_{t_{[1:n]}\in\mathcal{T}^n} f(w_{[1:n]}, t_{[1:n]})\cdot v$$

Can use the Viterbi algorithm, a dynamic programming algorithm that is $O(k^3 n)$, where k is the number of possible tags.

Oak Ridge National Laboratory — U.S. Department of Energy Office of Science — Homeland Security — PennState 1855 — ORAU — Oak Ridge Institute for Science and Education, Managed by ORAU for DOE