CERIAS

The Center for Education and Research in Information Assurance and Security

Using Syntactic Features for Phishing Detection

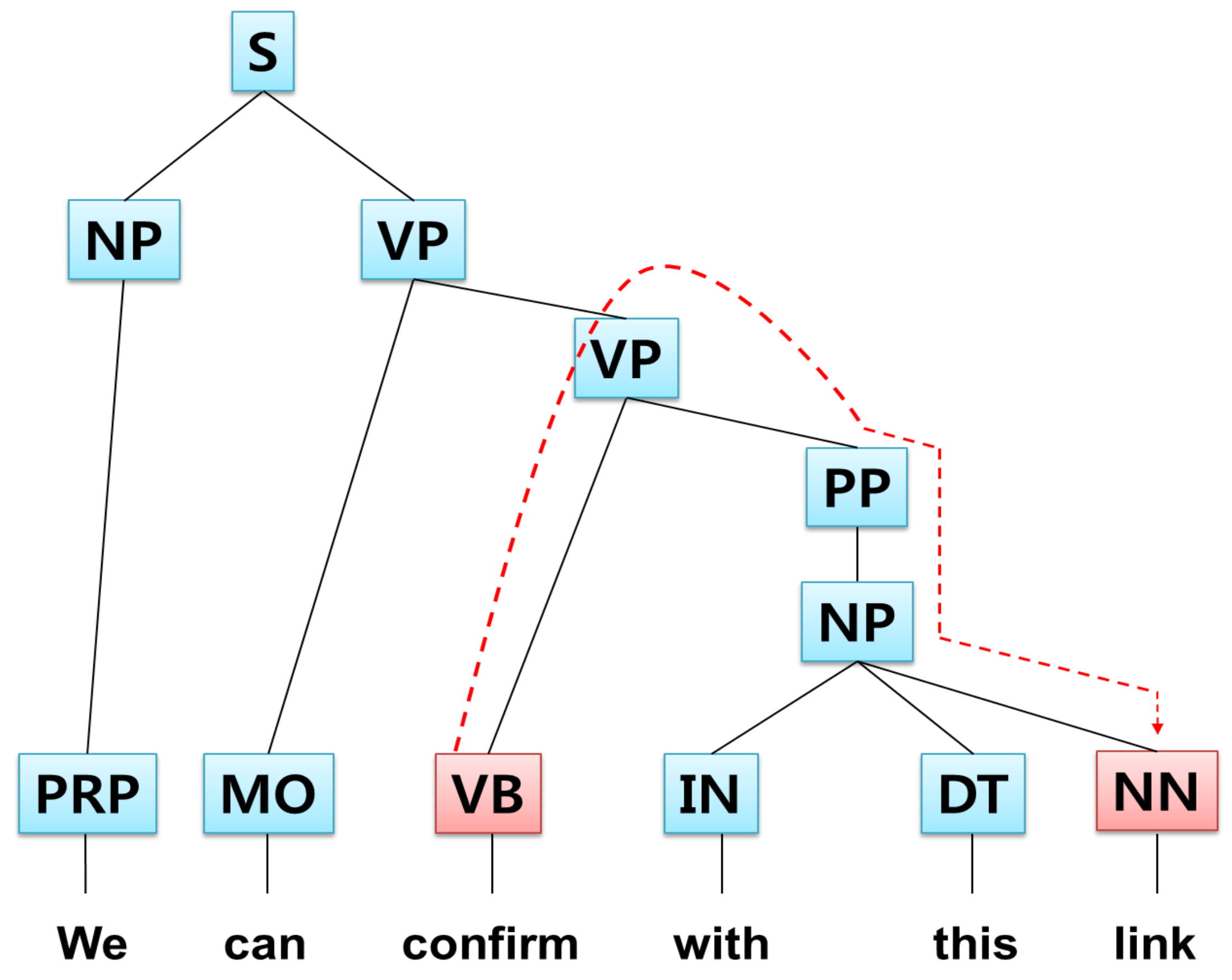
Students: Gilchan Park / Advisor: Julia M. Taylor

Abstract

The purpose of this research is to explore whether the syntactic structures and subjects and objects of verbs can be distinguishable features for phishing detection. To achieve the objective, we have conducted two series of experiments: the syntactic similarity for sentences, and the subject and object of verb comparison. The results of the experiments indicated that both features can be used for some verbs, but more work has to be done for others. The phishing corpora is comprised of old and up-to-date phishing emails, and the gap between them is over 10 years. To observe whether the pattern in phishing emails have changed over time with respect to subject and object of the verbs, we additionally compared between the two phishing corpora. The results showed us that most of subjects and objects were still identical, or similar from semantic perspective.

A. Parse Tree Path

Fig. 1. The parse tree path from '*confirm'* to '*link*' is represented as ↑VB↑VP↓PP↓NP↓NN



- Pare Tree Path - route from target verb to end node in syntactic tree
- **Test data**: phishing data (old emails in *2005*: # 2856, new emails in *2014*: # 25706)
  legitimate data (sample emails from Enron: # 3828)

TABLE I. THE COSINE SIMILARITY FOR PARSE TREE PATHS BETWEEN LEGITIMATE AND CORRESPONDING PHISHING DATA

Verb	Cosine similarity	
	Legitimate vs. Old Phishing	Legitimate vs. New Phishing
access	0.2923	0.2234
click	0.5865	0.6944
confirm	0.4388	0.5094
enter	0.5279	0.5503
follow	0.269	0.4184
protect	0.4196	0.5936
update	0.5547	0.6729
use	0.5789	0.7028

Results

- Most scores ranged between 40 to 70 percent. The values vary depending on the verbs.
- The difference in the parse tree paths was strongly affected by verbs themselves
- Not enough to be a distinguishable feature for phishing emails (for all verbs) because of the insufficient and inconsistent similarity scores

B. Subject and Object similarity

TABLE II. THE MOST FREQUENT SUBJECT OBJECT OF THE VERB *UPDATE*

Verb '*update*'	Subject	Old Phishing	**you** (75.58%) *That requires you to update the order Information.*
		New Phishing	**you** (71.84%) *You are required to update through the link below.*
		Legitimate	**you** (43.60%) *From there you will be able to update your email information securely.*
	Object	Old Phishing	**records** (47.79%) *Please update your records in maximum 24 hours.*
		New Phishing	**information** (25.93%) *Please update your information within 72 hours.*
		Legitimate	**profile** (17.24%) *If you're not signed in, you will need to do so before you can update your profile.*

TABLE III. THE COSINE SIMILARITY FOR SUBJECTS AND OBJECTS BETWEEN LEGITIMATE AND CORRESPONDING PHISHING DATA

Verb	Cosine similarity			
	Subject		Object	
	vs. Old	vs. New	vs. Old	vs. New
access	0.9868	0.544	0.0733	0.0241
click	0.9824	0.9652	0.9066	0.9003
confirm	0.2433	0.3402	0.0153	0.0513
enter	0.8712	0.883	0.227	0.2133
follow	0.6489	0.6555	0.22	0.3162
protect	0.001	0.0485	0.0724	0.1715
update	0.8769	0.8953	0.2316	0.4152
use	0.7364	0.8345	0.2372	0.4229

Results

- The most frequent subjects between phishing and legitimate were quite similar, However, the most frequent subjects in phishing emails were more dominant than those in legitimate emails
- The most frequent objects between phishing and legitimate were all different except for the verb *click*.. This is easily explainable as a number of things that are clickable is limited in normal life as well.
- The cosine similarity between the two phishing data: the results indicated that most verbs had similar subjects and objects between the two. Some exceptions appeared in objects (e.g. records, information), but they are in the same semantical domain.

PURDUE
UNIVERSITY