Poster: Syntactic Element Similarity for Phishing Detection

Gilchan Park
CIT
Purdue University
West Lafayette, IN USA
park550@purdue.edu

Julia M. Taylor
CIT & CERIAS
Purdue University
West Lafayette, IN USA
jtaylor1@purdue.edu

***Abstract*—This poster present the result of the comparison of the subject and object of verbs in their usage between phishing emails and legitimate emails. This research aims to investigate whether subjects and objects of verbs can be distinguishable features for phishing detection. This poster also reports the same comparison between old and up-to-date phishing emails to explore if patterns in phishing emails have changed over time. To accomplish the goal, we have conducted the experiments using two phishing corpora and a legitimate corpus. The results indicated that the feature can be used for some verbs, but more work has to be done for others.**

Keywords—phishing detection; subject; object; target verbs.

I. Introduction

The purpose of this poster is to report the comparison between phishing emails and legitimate emails, and between old phishing emails with up-to-date phishing emails in terms of subjects and objects of verb phrases for sentences. Our main interest was to investigate whether this feature could work as distinguishable characteristics for phishing emails and to observe whether the patterns in phishing emails have changed over time, with respect to subject and object of the verbs.

The dataset consists of old phishing emails collected in 2005 [1], up-to-date phishing emails reported in 2014 [2], and legitimate emails [3]. We have performed the experiments using the list of target verbs frequently appearing in phishing emails. In this poster, a subject is defined as the noun phrase which serves as the subject of a verb within a segment of a sentence, such as a clause. An object is referred as a noun or a pronoun which is the head of the syntactic object of the verb. For instance, in sentence *We need you to confirm your identity*. we define the subject of the verb *confirm* as *you* and the object of the verb *confirm* is *identity*.

For our purposes, we hypothesized that a verb might bring up different elements in a sentence depending on the intention of usage of the verb. For instance, scammers may use the word *update* in an attempt to gain personal information such as *update your account*, on the other hand, 'normal' users probably choose update for other purposes in email as well, such as in *I'll update the document*.

II. Prior Content based Phishing Detection Techniques

Some of content based phishing approaches have been proposed. For example, Hajgude and Ragha [4] and Xiang et al. [5] proposed phishing detection approaches using characteristics extracted from the contents. In particular, Verma et al. [6], Park & Taylor [7], Park [8] focused on analyzing the natural language features in text of the emails. These works analyzed phishing emails primarily based on keyword matching. This poster suggests an approach that goes beyond simple word comparison by expanding the scope of the analysis unit from words to sentence segments.

III. Experimental setup

The phishing data are composed of old phishing emails and up-to-date phishing emails, referred here as the Nazario corpus and the APWG corpus respectively. The Nazario corpus was taken from a publicly available collection of phishing emails[1], and the APWG corpus was constructed from the emails provided by Anti-Phishing Working Group [2]. The APWG corpus are reported emails by users to the group. In this experiment we used phishing emails reported in September 2014. The legitimate emails are from the public Enron email set by the CALO Project[3].

We preprocessed the data sets to remove duplicates, non-English emails, and unnecessary information, such as headers, forwarded text, etc. After cleaning up the data, the size of corpora was 2,746 emails for Nazario, 30,375 emails for APWG, and 237,440 emails for the legitimate corpus. The target verbs were the most frequent verbs appeared in both phishing corpora in common: *access, change, click, confirm, contact, enter, follow, need, pay, protect, require, update, use, verify, visit.*

To find the subject and the object of a verb in a sentence segment, we adopted the Stanford Parser (version 3.4.1) that provides the Stanford typed dependencies representation (SD) [9]. SD represents the simple description of grammatical relations in a sentence as binary relations. The binary relations indicate the grammatical relations between the dominating constituent such as a verb and the dependent or dominated constituent such as a subject or an object of the verb.

A. Similarity between Legitimate and Phishing Corpora

We applied the cosine similarity measurement to compare the subjects and objects between the two data. Two vectors for a verb contained the occurrences of subjects and objects of the verb in each dataset as the values of the vectors. The results of cosine similarities are shown in TABLE I.

This research was partially funded by a seed grant from the College of Technology of Purdue University

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

Verb	Cosine similarity			
	Legit vs. Nazario		**Legit vs. APWG**	
	Subject	***Object***	***Subject***	***Object***
access	0.9258	0.1434	0.7062	0.3881
change	0.6881	0.3348	0.745	0.4079
click	0.9369	0.6929	0.9037	0.7665
confirm	0.5241	0.1188	0.6548	0.2326
contact	0.8574	0.3083	0.9421	0.5034
enter	0.6157	0.4021	0.6948	0.2615
follow	0.4706	0.6497	0.4973	0.7396
need	0.8225	0.5579	0.8653	0.7416
pay	0.4556	0.1371	0.49	0.3641
protect	0.4937	0.1232	0.4082	0.2979
require	0.366	0.4522	0.558	0.2761
update	0.648	0.2408	0.6683	0.3679
use	0.6491	0.1787	0.8158	0.4706
verify	0.5395	0.2617	0.7434	0.1734
visit	0.7859	0.1428	0.8659	0.7226
Avg.	**0.6519**	**0.3163**	**0.7039**	**0.4476**

In overall, the subjects showed more similar than the objects. Looking at individual results, the similarity seemed to depend on verbs themselves. The verbs *follow, pay, protect, require* had a much lower score in subject similarity than the other verbs, and the verbs *click, follow, need* had a significantly greater score in object similarity than the other verbs.

B. Similarity between the two Phishing Samples

The TABLE II reports the cosine similarities between the two phishing corpora. All verbs had quite similar subjects. The objects were not as analogous as the subjects, but the similarity scores were relatively higher than those in legitimate corpus.

TABLE II. THE COSINE SIMILARITY FOR SUBJECTS AND OBJECTS BETWEEN NAZARIO AND APWG

Verb	Cosine similarity	
	Subject	***Object***
access	0.7396	0.2978
change	0.9218	0.6439
click	0.9103	0.8968
confirm	0.9607	0.8255
contact	0.8904	0.8992
enter	0.9893	0.5759
follow	0.8947	0.7348
need	0.6995	0.7547
pay	0.8149	0.2575
protect	0.6149	0.6292
require	0.7933	0.5116
update	0.9807	0.5127
use	0.9146	0.2323
verify	0.6274	0.8192
visit	0.9492	0.3497
Avg.	**0.8468**	**0.5961**

IV. Conclusion

In this poster, we presented the difference in subjects and objects of target verbs between phishing emails and legitimate emails and between the two phishing email sets. The results between phishing and legitimate showed that the similarities in the subjects were greater than those in objects for most verbs. A possible explanation is that objects play a bigger role in delivering the sender's intent, thus significantly limiting the domain. The similarity scores were not consistent, and seemed to depend on verbs themselves rather than a clear-cut differentiation between phishing or legitimate categories. The comparison between the two phishing sets indicated that the subjects were still quite similar, but the objects were not as similar as the subjects. However, the different objects still had similar meanings.

The selected target verbs also frequently appeared in the legitimate emails. If we simply adopt a phishing detection method based on keyword matching, it will cause false positives. The suggested approach was the first step to go beyond simple word comparison by expanding the analysis unit from words to sentence segments. Our future research is to find patterns in email texts in terms of word meanings, and cluster them into semantic domains. We expect this work will be able to handle the syntactically different, but semantically identical or similar words to identify the intention of email, and produce features to be generalized.

REFERENCES

[1] Nazario, J. (2005). The online phishing corpus, Available from: http://monkey.org/~jose/wiki/doku.php

[2] Anti-Phishing Working Group. APWG Phishing Archive.

[3] Yorke-Smith, N., Saadati, S., Myers, K. & Morley, D. (2012). The design of a proactive personal agent for task management, International Journal on Artificial Intelligence Tools 21:1.

[4] Hajgude, J., & Ragha, L. (2012). Phish mail guard: Phishing mail detection technique by using textual and URL analysis. In Information and Communication Technologies (WICT), 2012 World Congress on IEEE, 297-302.

[5] Xiang, G., Hong, J., Rose, C.P,. & Cranor, L. (2011). CANTINA+: A feature-rich machine learning framework for detecting phishing websites, ACM Transactions on Information and System Security (TISSEC) 14:2.

[6] Verma, R., Shashidhar, N., & Hossain, N. (2012). Detecting phishing emails the natural language way, Computer Security–ESORICS 2012.

[7] Park, G., & Taylor, J. M. (2013). Towards text-based phishing detection, Proceedings of SDPS Conference, Sao Paolo, Brazil.

[8] Park, G. (2013). Text-Based Phishing Detection Using a Simulation Model. Masters' Thesis, Computer and Information Technology, Purdue University, W. Lafayette, IN, USA.

[9] De Marneffe, M. C., & Manning, C. D. (2008). The Stanford typed dependencies representation. In Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation (pp. 1-8). Association for Computational Linguistics.