Understanding Multi-lingual Threat Intelligence for AI based Cyber-defense Systems

Priyanka Ranade, Sudip Mittal and Anupam Joshi
University of Maryland, Baltimore County, Baltimore, MD 21250, USA
Email: {gy63575, smittal1, joshi}@umbc.edu

Abstract

Information across political, cultural, and geographical boundaries is widely communicated over a global Internet. Today we have a multilingual Internet where people converse in languages like English, Mandarin, Russian, Hindi, etc [1]. Cyber threats in particular, originate from and are mitigated over a broad range of geographic regions. Although cybersecurity web data is vastly available on the web, it is disparate among major natural languages, decreasing interoperability on a multilingual level. The vast geographic distribution of cyber attacks increases the difficulty of employing strong cyber risk management across organizations worldwide.

Cybersecurity actors, both attackers and defenders, converse over social media, blogs, dark web vulnerability markets,etc in diverse languages. These *non-traditional* sources are becoming an important asset for threat intelligence mining and many times are first to receive the latest intelligence about vulnerabilities, exploits, and threats [4]. These sources are prime tools for dissemination of integral threat intelligence data, ranging from political factors such as the international origination and intention behind attacks, to technical factors such as sources of new software vulnerabilities and exploits.

The multilingual nature of these non-traditional sources is a potential hindrance for cyber-defense professionals, as they might be limited by their knowledge of different languages. Despite this significant issue, the role of language in sourcing cyber threat has been under explored. The security industry is heavily contingent upon the security analyst's ability in using specialized experience to reason over the disparate pieces of intelligence data available on the web, in order to discover potential threats and attacks.

Cybersecurity web data has induced the concurrent use of artificial intelligence based cyber-defense systems to help analysts extract relevant pieces of information that may constitute an attack. These systems need an ability to process, understand and then include multiple languages to keep up to date with the most current threat intelligence. A multilingual Internet needs a multilingual approach to cybersecurity.

In this work, we propose a multilingual processing system that harnesses critical disparate cybersecurity data derived from various natural languages to address the international nature of cyber attacks and assist in defensive cyber operations. This system creates a representation for cybersecurity data present in the English and Russian languages. We investigate semantic representation of multiple languages with a cybersecurity corpus from Twitter[1] about cybersecurity threats and vulnerabilities in two natural languages, English and Russian.

We first collect data through the Twitter streaming API[2] based on cybersecurity keywords popular in the English and Russian languages. The keywords stemmed from technical terms such as "overflow" to more political terms such as "cyberwarfare". The data is dynamically stored and separated by language in MongoDB[3], an open source NoSql Database[4]. We then, produce separate vector models out of the Russian and English tweets, through Word2Vec [3, 2]. Word2vec is a predictive model for learning word embeddings from raw text. It proposes two models, the Continuous Bag-of-Words model (CBOW) and the Skip-Gram model. In our overall architecture, we hope to find similar foreign security terms from context, therefore, we employ the CBOW method which takes multiple words as targets, and utilizes words surrounding the target input to predict similar words as one output. We then evaluated the word embeddings by dividing each vocabulary into subsets containing synonyms of words through Wordnet [5] a lexical database for the English (and supported) languages and the Wordnet Synset library[5], which groups words into sets of synonyms. The synonyms were then checked against each space's vocabulary in order to measure the accuracy of each model. The subsets that we created were then placed into four manually curated logical sets based upon modern SOC technology strategy in sourcing attacks (technical: indicators of specific malware, viruses, etc), strategic (tweets pertaining

[1] https://www.twitter.com/
[2] https://developer.twitter.com/en/docs
[3] https://www.mongodb.com/
[4] https://www.mongodb.com/nosql-explained
[5] https://wordnet.princeton.edu/

to attacker architectural methodology), strategic (organizational/political information regarding attacks), and anticipated threat (details of potential cyber attacks).

Bilingual vector spaces present many potential applications that can be deployed in real world SOC settings, and future implications go beyond the technical, by also ultimately impacting organizations on a cyber policy level as well. The synsets, as well as the logical sets mentioned above, can be utilized in future tasks of transferring knowledge from one language space, in order to generalize common and similar cybersecurity terms across lesser known languages. These learned embeddings in various languages can be utilized to categorize an international response system at different levels, each level indicating consequences from organizational data disruptions. The consequences and response mechanisms will differ depending on the societal and economic structure pertaining to the regional natural language present.

References

[1] The U.S. Census Bureau. Internet world stats. https://www.internetworldstats.com/, Dec 2017.

[2] T Mikolov and J Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 2013.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *preprint arXiv:1301.3781*, 2013.

[4] The Register. Most vulnerabilities first blabbed about online or on the dark web. https://www.theregister.co.uk/2017/06/08/vuln_disclosure_lag/, Jun 2017.

[5] Princeton University. Wordnet: An electronic lexical database. 2010.