Using Deep Neural Networks to Translate Multi-lingual Threat Intelligence

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

***Abstract*—The multilingual nature of the Internet increases complications in the cybersecurity community's ongoing efforts to strategically mine threat intelligence from OSINT data on the web. OSINT sources such as social media, blogs, and dark web vulnerability markets exist in diverse languages and hinder security analysts, who are unable to draw conclusions from intelligence in languages they don't understand. Although third party translation engines are growing stronger, they are unsuited for private security environments. First, sensitive intelligence is not a permitted input to third party engines due to privacy and confidentiality policies. In addition, third party engines produce generalized translations that tend to lack exclusive cybersecurity terminology. In this paper, we address these issues and describe our system that enables threat intelligence understanding across unfamiliar languages. We create a neural network based system that takes in cybersecurity data in a different language and outputs the respective English translation. The English translation can then be understood by an analyst, and can also serve as input to an AI based cyber-defense system that can take mitigative action. As a proof of concept, we have created a pipeline which takes Russian threats and generates its corresponding English, RDF, and vectorized representations. Our network optimizes translations on specifically, cybersecurity data.**

I. Introduction

Information across political, cultural, and geographical boundaries is widely communicated over a global Internet. Today, we have a multilingual Internet where people converse in a variety of languages like English, Mandarin, Russian, Hindi, etc. [4]. Cyber threats in particular, originate from and are mitigated over a broad range of geographic regions. Although a significant amount cybersecurity web data is available, it is spread among major natural languages, decreasing interoperability between multilingual systems. This creates difficulty in employing strong cyber risk management across organizations worldwide. Specifically, amongst state actors or major criminal networks, it is likely that the threat information is in a language other than the language of the analyst.

Intelligence gathering spans an expansive geographic distribution. As a result, cybersecurity actors, both attackers and defenders, converse over *non-traditional sources* such as social media, blogs, dark web vulnerability markets, etc. in diverse languages. These non-traditional sources are becoming an important asset for threat intelligence mining [32] and many times are first to receive the latest intelligence about vulnerabilities, exploits, and threats [31]. 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 addressing cyber threats has been under explored. Multilingual understanding, adds to the many challenges security analysts continue to encounter. The security industry is heavily dependent 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.

The abundance of cybersecurity web data has led to the use of AI/NLP based cyber-defense systems to help analysts extract relevant pieces of information that may constitute an attack. These systems need the ability to process multiple languages to keep up to date with the most current threat intelligence. A multilingual Internet needs a multilingual approach to cybersecurity.

While modern cyber defense systems have the ability to reason over disparate pieces of threat intelligence data on the web, we hope to create a defensive system that also understands various languages, by using the English language as a baseline. In our previous work, we developed *CyberTwitter* and *Cyber-All-Intel* [24], [25], systems that mine threat intelligence data from various sources, and automatically issues cybersecurity vulnerability alerts to users. This work extends these cyber-defense systems to a wider spectrum of potential threats, by mining threat intelligence data in a multitude of languages. These systems typically produce "cyber terminology representations" [24], [25] to categorize threat-related words, but only learn representations for English. Consequently, if a certain threat is not gathered under a specific language, the system will not have a representation for it, even if it is a known threat in a different language. We use our multilingual threat intelligence system to align cyber terminology representations of different languages, expanding monitoring capabilities across the globe.

In this work, we create a multilingual translation system that harnesses critical cybersecurity data derived from various natural languages to address the international nature of cyber attacks and assist in defensive cyber operations. Our system optimizes translations particularly for cybersecurity data. Specifically, we investigate semantic representation of multiple languages with a corpus from Twitter, including threats and vulnerabilities in two languages, English and Russian. We build models to relate the vector space representations in the two languages to translate threat from Russian to English.

Our overall use case (see Figure 1), utilizes embeddings created from Russian and English threat intelligence data. The embeddings help us understand security terms in Russian, by

978-1-5386-7848-0/18/$31.00 ©2018 IEEE

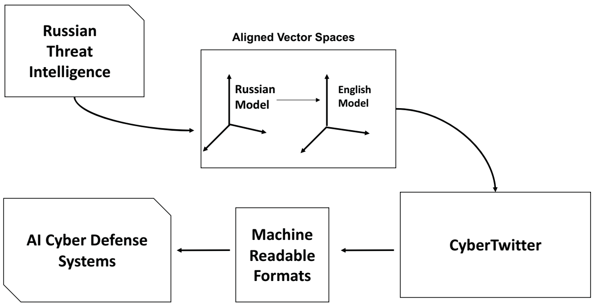


Fig. 1. Multilingual Threat Intelligence Platform

aligning semantically similar Russian cyber terms with their English counterparts. The system first begins by gathering relevant Russian threat intelligence data from sources such as Twitter. The data is then assimilated into a vector representation in order to bring semantically similar terms together [21]. The data is then fed into CyberTwitter, which converts the English representation of the Russian data into a machine understandable format defined using our UCO Ontology [35] in OWL. This helps cyber-defense systems gain intelligence about threats mentioned in the Russian text. The acquired intelligence is then fed into an AI-based cyber defense system that generates conclusions from a cumulation of aggregated threat intelligence data.

An issue with directly converting threats in foreign languages to machine readable formats, is removing the security analyst from the threat inspection process. Providing analysts with raw translations help them reason over and expand upon a new landscape of threats and vulnerabilities. Our system aims to therefore, serve as an augmentation system that helps analysts divert full attention on their primary roles of analyzing and piecing in novel threat information.

The rest of the paper is organized as follows - Section II contains the related work. We discuss our intelligence translation framework in Section III. We present our experimental results and evaluations in Section IV. We discuss how we use the translation system with an AI based cyber-defense system in Section V. We conclude in Section VI.

II. Related Work

In this section, we present related work on the vector space model uses, neural machine translation, AI-based cybersecurity systems, and cybersecurity understanding across different languages.

A. Text Analysis for domain specific tasks

Text analytics has been utilized in areas such as information retrieval [17], machine translation [14], and topic detection [19]. These areas are especially useful for domain specific tasks such as cybersecurity.

Vector Space Models: Vector Space Models, or word embeddings, have been used in Natural Language Processing. Words are embedded in a continuous vector space such that, words that appear in the same contexts are semantically related. One method that generates embeddings based on word co-occurrence is word2vec [21], [20]. Mikolov et al. [20], showed that proportional analogies can be solved by finding the vector closest to the hypothetical vector. Embeddings have also been utilized in other areas such as word sense disambiguation [3], semantic search [34], and discovering inter-linguistic relations in machine translation studies [23].

Wordnet [22] is a human curated lexical database that groups together synonyms in the English language. Many other versions of Wordnet have been produced, such as Arabic-WordNet and ChineseWordnet [28]. These lexical databases are often times used to aid lexical and term alignment.

Term Alignment in Vector Spaces: Analogical Relationships are often times utilized to aid term alignment in vector spaces. Term alignment is known as statistically finding correspondences between words in different groups [5]. Plas et al. [18] utilize automatic world alignment to find translations from Dutch to one or more target languages. Similarly, Brown et al. [29] aligned sentences with their translations in two parallel corpora, consisting of French and English. Yang et al. [26] show how the pattern of the context from word embeddings help to align similar word pairs in other languages. Piantra et al. [10] created MultiWordnet, an aligned multilingual database curated to produce an Italian Wordnet, by aligning synonyms in Italian to EuroWordNet. Niemann et al. [9] aligned WordNet synonym sets and Wikipedia articles to group article topics based on synonyms.

B. Neural Machine Translation

Word embeddings have aided in a diversity of machine translation tasks. Neural machine translation typically operates through the encoder-decoder-attention architecture [7]. More recently, bilingual word distributions have been trained using unsupervised methods such as Latent Dirichlet allocation (LDA) and Latent semantic analysis (LSA) to aid machine neural translation [15]. Lample et al. [13] trained word embeddings from monolingual data and utilized external and internal vectors as input for the network utilized to train unfamiliar instances of words. In terms of semantic translation tasks, Hill et al. [11] show that translation-based embeddings work better in applications that require concepts organized according to similarity.

C. Cybersecurity understanding across multiple languages

Cybersecurity terminology definitions differ across cultures and languages. The Department of Homeland Security started developing multilingual resources, to help link cybersecurity understanding across international governments [6]. Klavens et al. [16] outlines the importance of linguistics in the domain of security and claims language analysis propels understanding of communication between cyber-crime activist groups, filtering relevant data collection, and understanding the intention behind the words.

D. AI-Based Cyber Defense Systems

The use of social media in threat intelligence mining, provides a new interface between the public and the Intelligence Community. Twitter data in particular, is seen as a reliable OISNT resource due to its real time nature during high impact events, such as terrorist attacks [1]. Mittal et al. [24] developed CyberTwitter, a threat intelligence framework that utilizes twitter data to automatically issue security vulnerability alerts to users. Similarly, the Cyber-All-Intel system collects OISNT data, stores it in a cybersecurity corpus, and utilizes word vectors for cybersecurity term similarity searches [25].

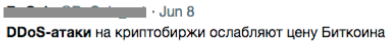
III. Intelligence Translation Architecture

In this section, we describe our data collection methods, vector space generation, alignment techniques and neural machine translation framework. We first create a multilingual cybersecurity corpus that contains tweets about threats and vulnerabilities in various languages. In this paper, to create a proof of concept, we focus on English and Russian. Using the collected corpus,we then produce English and Russian vector embeddings. Once we create the embeddings, we align both vector spaces utilizing an alignment database. Once the spaces are aligned, we are able to undertake semantic translation of Russian cyber threats and vulnerabilities to English.

A. Creating a Multi-Lingual Cybersecurity Corpus

We collect data through the Twitter streaming API using cybersecurity keywords: *XSS, CVE, spam, malware, data, attacker, DNS, DDOS, code, ciphertext, cryptography, hacked, overflow, breach, sniffer, buffer, firewall, domain, hijacking, checksum, virus, vulnerability, arbitrary, protocol,* etc. These keywords were suggested by multilingual cybersecurity domain experts [27] and various security analysts. We use the Twitter API language capabilities to detect "tweet language" through a flag (en=English, ru=Russian). Setting this flag provides us the ability to collect data in both languages. The data is stored and separated by language in MongoDB.

Collecting data using these keywords, gives us a direct interface to Russian cyber colloquialisms. For example, the tweet depicted in Figure 2, reveals a regional-specific DDoS attack, to threat analysts outside of Russia.



[redacted] · Jun 8

DDoS-атаки на криптобиржи ослабляют цену Биткоина

Fig. 2. Sample tweet from a Russian corporation on crypto DDoS attacks that translates to "DDos attacks on cryptocurrency weaken the price of Bitcon."

B. Embedding Generation

We generate English and Russian word embeddings using Word2Vec [21], [20]. We created two separate vector space models from the English tweets and the Russian tweets.

In our system, these models are used for semantic translation of Russian cybersecurity text to English (see Section III-C1). Words in the embedding space are *semantically similar* if grouped together around the same neighborhood. For example, in our English model, words like, malware, virus, trojan, etc. will be clustered together. Figure 3, depicts a 20th iteration training snapshot, of Russian words that start appearing near "DDoS", a type of cybersecurity attack.

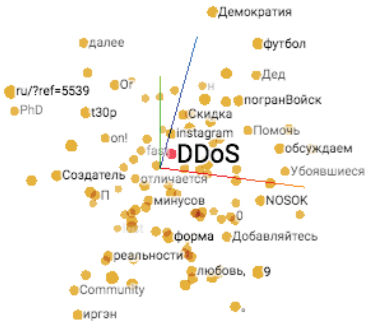


Fig. 3. Russian Embedding Training Snapshot around the word "DDOS". Neighboring words include community and proxy.

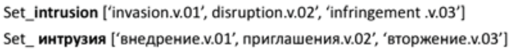
C. Intelligence Translation Framework

In this section, we describe our intelligence translation framework that creates many-many mappings among cybersecurity terms in Russian and English. We use the word embeddings produced in Section III-B, and create an alignment database, later used as the dataset for training our neural network described in Section III-C2.

In a high level usecase, a cyber-defense system like CyberTwitter [24], will take as its input Russian threat intelligence and create machine readable threat intelligence. This scenario is discussed in detail in Section V.

1) Creating an Alignment Database: In order to create relationships between English and Russian cybersecurity words, we created a dataset to align the English and Russian vector embeddings. An alignment in our system means creating true positive mappings of Russian cybersecurity terms to their English counterparts. We derived *cybersecurity synsets* for the Russian and English vocabulary embeddings, created in section III-B. These cybersecurity synsets include contextually similar words to each vocabulary word in the Russian and English vector spaces. We emphasize that, when we say contextually similar words, we bring together cybersecurity terms in the same word sense. The lexical database Wordnet [22], groups similar words into sets of synonyms called "synsets". WordNet does not support the Russian language. We found a similar lexical database called Wiki-Ru-Wordnet[30], specifically for the Russian language. We utilize the English synsets provided by WordNet, and the Russian synsets provided by Wiki-Ru-Wordnet to create our cybersecurity synsets. An example of a cyber synset we derived is shown in Figure 4.

We tasked three native Russian speakers, who served as annotators, to manually verify the quality of cybersecurity



Set_**intrusion** ['invasion.v.01', disruption.v.02', 'infringement .v.03']
Set_ **интрузия** ['внедрение.v.01', приглашения.v.02', 'вторжение.v.03']

Fig. 4. English and Russian Synset Examples for "intrusion".

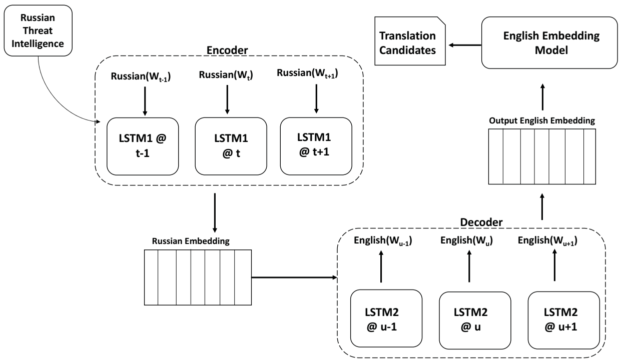


Fig. 5. Intelligence Translation Architecture

synsets produced. We use the Cohen's kappa to compute the inter-annotator agreement, and keep only those cybersecurity synsets that scored higher than 0.66. The annotators confirmed that the synsets in Wordnet, and the synsets in Wiki-Ru-Wordnet, were not only similar on a translation level, but also semantically similar in a cultural context.

We use these cybersecurity synsets to then create a neural network, that given an input Russian embeddings, outputs its English equivalent.

2) Intelligence Translation Network: Our intelligence translation architecture is shown in Figure 5. We implement a encoder-decoder network, which is a dual Recurrent Neural Network (RNN). The encoder serves as an input RNN and the decoder serves as the output. The encoder-decoder architecture projects the input Russian word to be translated into the English embedding space, by returning words with a representation closest in the English vocabulary.

The encoder-decoder network is implemented using a "sequence-to-sequence architecture" [8], [33], [12]. The encoder-decoder network is able to process past and future words in a sequence, and is also able to map an input sequence to an output sequence of a different length.

The model uses word embeddings produced in section III-B. We initialize Russian embeddings as the input in the encoder state, and the English embeddings as the output of the decoder state. We utilize the cybersecurity synsets, from section III-C1, as the training set. Our hyperparameters were set as, batch size = 64, epochs = 100, latent dimentionality = 256, sample number = 10,000.

The encoder and decoder utilize a Long Short Term Memory (LSTM) cell [8]. In the hidden layer, we have one dense layer, with a softmax activation function, which allows the model to learn a mapping from the Russian vector representation, to the English vector representation. The encoder, takes in Russian words and maps them to their respective vector representations. The decoder then, creates a translation of the input word and generates its predicted aligned semantic English embedding. We discuss the accuracy for this model in Section IV.

IV. Evaluation

In this section, we describe our experimental setup and evaluate our intelligence translation system. We first measure our translation precision through BLEU and accuracy metrics. Later, we compare our system against other commercial translation services.

A. Accuracy and BLEU score

We first evaluate our encoder-decoder architecture through an accuracy metric and a BLEU (Bilingual Evaluation Understudy) score (see Table 1). The accuracy metric computes the percentage of times that predictions match labels. BLEU scores, are standard metrics for evaluating a generated translation to a reference word [12]. An accuracy above "60%", perplexity under "6", and a BLEU Score between "15 and 36" is considered robust [12].

Measure	**Value**
Accuracy	97.22%
Perplexity	4.07
BLEU Score	28.4

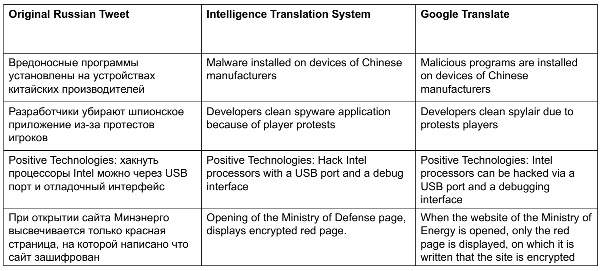
TABLE I
Evaluation Metrics

B. Measuring against Commercial Systems

We measure the precision of our translations by checking a randomly generated sample of the output against Google Translate[1]. We proved that our system produces more effective translations for the security domain. We extracted 1000 randomly selected tweet translations and compared the output against the Google Translate API. We check our translations against the ones provided by Google Translate, both syntactically and semantically. On evaluating 1000 random samples, there was a 64.3% syntactic correlation between the two systems.

We further evaluated the 357 samples that were not syntactically equivalent and tasked two security analysts to manually evaluate the semantic meanings of the translation outputs. We found that of the 357 outputs, 349 were semantically similar, but not syntactically similar to the Google translation, showing 97% semantic relevance. The annotators concluded our translations are preferable through a security perspective, in that they proliferate terms unique to the security industry. The commercial translation services are generalized while our system is domain specific. These security specific translations can be attributed to the architecture of our model, that utilizes a specialized aligned database made with relevant cybersecurity mappings. Examples of unequal but semantically similar translations in our system and Google Translate are listed in figure 6. In example 1, "malware" registers more with a security analyst than "malicious programs". In example 2, the

[1] https://translate.google.com



Original Russian Tweet	Intelligence Translation System	Google Translate
Вредоносные программы установлены на устройствах китайских производителей	Malware installed on devices of Chinese manufacturers	Malicious programs are installed on devices of Chinese manufacturers
Разработчики убирают шпионское приложение из-за протестов игроков	Developers clean spyware application because of player protests	Developers clean spylair due to protests players
Positive Technologies: хакнуть процессоры Intel можно через USB порт и отладочный интерфейс	Positive Technologies: Hack Intel processors with a USB port and a debug interface	Positive Technologies: Intel processors can be hacked via a USB port and a debugging interface
При открытии сайта Минэнерго высвечивается только красная страница, на которой написано что сайт зашифрован	Opening of the Ministry of Defense page, displays encrypted red page.	When the website of the Ministry of Energy is opened, only the red page is displayed, on which it is written that the site is encrypted

Fig. 6. Tweet Translation Samples.

Google Translate system translated the relevant Russian text as “spylair”, while our system gives the correct translation as “spyware”. These are clear instances in which our translation will provide more relevant and direct intelligence for a security professional.

Another benefit that our system provides is that it can run independently in secluded operational settings. A security analyst may not be able to input their sensitive data into third party platforms due to privacy, security, and confidentiality policies.

V. Use by Cyber-Defense Systems

Web based unstructured, textual sources such as Twitter, Reddit, blogs, dark web forums, etc. provide a rich multilingual source of information about cyber threats and attacks. In addition to providing details of existing attacks, such sources (especially the dark web) can serve as advance indicators of attacks in terms of discussions around newly discovered vulnerabilities. This information is available in textual sources traditionally associated with Open Sources Intelligence (OSINT), as well as in data that is present in hidden sources like dark web vulnerability markets.

The intelligence translation system that we discuss in Section III will help us automate this process by taking data from a variety of multilingual sources, extracting, representing and integrating the knowledge present in it as embeddings and knowledge graphs, and then use the resulting artificial intelligence systems to provide actionable insights to SoC professionals. Figure 7 showcases our pipeline, which takes in Russian threat intelligence and stores it in as a *VKG structure* [25].

Two such systems that we have developed in the past are CyberTwitter [24] and Cyber-All-Intel [25]. The systems store threat intelligence in a knowledge representation that can be used by AI based cyber-defense systems (See Figure 7). Such systems generally have a knowledge representation engine, a reasoning engine, and few applications like an alert generation system, recommender system, query processing system, etc.

The knowledge representation system, converts input threat intelligence (usually in a textual format) into a machine readable format. In our system we represent it in RDF[2], with cybersecurity domain knowledge provided by the Unified Cybersecurity Ontology (UCO) [35]. The intelligence ontology [24] provides information about the intelligence domain. We also include specific conceptual embeddings for security concepts in our threat representation format [25]. The knowledge reasoning part of the system provides domain specific reasoning capability generally encoded as logical rules by a domain expert. The applications and the reasoning engine generally use the machine readable representation to provide specific functionality. Figure 7 also provides the graph structure for the translated English intelligence: “*URL Command Injection Remote Code Execution Vulnerability in Microsoft Skype*”. Figure 8 provides the RDF representation for the same intelligence.

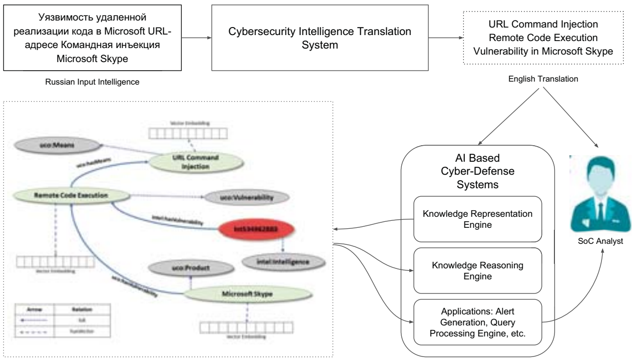

Уязвимость удаленной реализации кода в Microsoft URL-адресе Командная инъекция Microsoft Skype
Russian Input Intelligence
Cybersecurity Intelligence Translation System
URL Command Injection Remote Code Execution Vulnerability in Microsoft Skype
English Translation
uco:Means
URL Command Injection
uco:hasMeans
Remote Code Execution
uco:Vulnerability
intel:hasVulnerability
Int534962883
intel:intelligence
uco:Product
uco:hasVulnerability
Microsoft Skype
Vector Embedding
Vector Embedding
Vector Embedding
Arrow
Relation
AI Based Cyber-Defense Systems
Knowledge Representation Engine
Knowledge Reasoning Engine
Applications: Alert Generation, Query Processing Engine, etc.
SoC Analyst

Fig. 7. Using the intelligence translation system with an AI based cyber defense system.

```
@prefix uco: <http://accl.umbc.edu/ns/ontology/uco#> .
@prefix intel: <http://accl.umbc.edu/ns/ontology/intelligence#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix dbp: <http://dbpedia.org/resource#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .

<Int534962883> a intel:Intelligence ;
intel:hasVulnerability <remote_code_execution> ;

<command_injection> a uco:Means .

<Microsoft_Skype> a uco:Product ;
uco:hasVulnerability <remote_code_execution> ;
owl:sameAs dbp:Skype .

<remote_code_execution> a uco:Vulnerability ;
uco:affectsProduct <Microsoft_Skype> ;
uco:hasMeans <command_injection> ;
owl:sameAs dbp:remote_code_execution .
```

Fig. 8. RDF for textual input “URL Command Injection Remote Code Execution Vulnerability in Microsoft Skype”. Also, $owl : sameAs$ property has been used to augment the data using an external source ‘DBpedia’ [2].

[2] https://www.w3.org/RDF/

VI. Conclusion & Future Work

In this paper, we described the design, implementation, and evaluation of a multilingual threat intelligence translation system. The system uses Russian and English word embeddings created from cybersecurity data, an aligned cyber term database, and a LSTM based neural machine translation architecture, to translate cybersecurity text from Russian to English. With the help of Russian speaking cyber analysts, we created an alignment database by generating synonyms for the Russian and English corpus vocabularies, along with their respective translated Russian and English words. We utilize this database in neural machine translation, where we use an encoder-decoder architecture to map unfamiliar Russian cyber inputs to their English counterparts. We show that our model not only has high syntactic correlation to third party translation systems, but also registers prevalent cybersecurity terms in translation better than third party engines. We extend third party translation systems by creating a domain specific model that can provide more pertinent intelligence for an analyst. Our system can be utilized in private operational settings that do not permit the use of third party applications when dealing with sensitive intelligence data.

A weakness of our system, is the requirement of a cybersecurity rich alignment to train the model. Although we derived a Russian and English cybersecurity synonym sets in this proof of concept, it is an expensive task that will take dispersed effort across the linguistic and security communities, to derive across many other languages. In order to create more mappings for cyber terms across other languages like, Mandarin, Cantonese, Portuguese, Arabic, Hindi, etc. future research can include creation of multilingual cyber alignment databases. We can also consider transferring knowledge from languages with an abundance of intelligence to other unknown languages with no or few alignments. We expect that aligned cyber embeddings across many languages can promote international incident response collaboration.

Acknowledgement

The work was partially supported by a gift from IBM Research, USA and the Undergraduate Research Award from the University of Maryland, Baltimore County.

References

[1] Ponnurangam Kumaraguru Aditi Gupta. Credibility ranking of tweets during high impact events. 2012.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[3] Amir Bakarov. Improving word representations via global context and multiple word prototypes. 2018.

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

[5] R.Priyanga C.Sundar. Mining words and targets using alignment model. 2016.

[6] DHS. Stop.think.connect. multilingual resources. https://www.dhs.gov/stopthinkconnect-multilingual-resources/, 2015.

[7] Yoshua Bengio Dzmitry Bahdanau, Kyunghyun Cho. Neural machine translation by jointly learning to align and translate. 2014.

[8] Yoshua Bengio Dzmitry Bahdanau, KyungHyun Cho. Neural machine translation by jointly learning to align and translate. 2015.

[9] Jorg Tiedemann Elisabeth Niemann. The peoples web meets linguistic knowledge: Automatic sense alignment of wikipedia and wordnet. 2006.

[10] Christian Girardi Emanuele Pianta, Luisa Bentivogli. Developing an aligned multilingual database. 2002.

[11] Sebastien Jean Coline Devin Yoshua Bengio Felix Hill, Kyunghyun Cho. Embedding word similarity with neural machine translation. 2014.

[12] Yuntian Deng Jean Senellart Alexander M. Rush Guillaume Klein, Yoon Kim. Opennmt: Open-source toolkit for neural machine translation. 2017.

[13] Ludovic Denoyer Marc'Aurelio Ranzato Guillaume Lample, Alexis Conneau. Unsupervised machine translation using monolingual corpora only. 2017.

[14] Ashok C. Popat Moshe Dubiner Jakob Uszkoreit, Jay M. Ponte. Large scale parallel document mining for machine translation. 2010.

[15] Chengqing Zong Jiajun Zhang. Bridging neural machine translation and bilingual dictionaries. 2016.

[16] Judith L. Klavans. Cybersecurity - whats language got to do with it? 2015.

[17] Ray R. Larson. Introduction to information retrieval. 2009.

[18] Jorg Tiedemann Lonneke van der Plas. Finding synonyms using automatic word alignment and measures of distributional similarity. 2006.

[19] Claudio Schifanella Mario Cataldi, Luigi Di Caro. Emerging topic detection on twitter based on temporal and social terms evaluation. 2010.

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

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

[22] George A Miller. Wordnet: a lexical database for english. 1995.

[23] Piyush Kedia Pushpak Bhattacharyya Mitesh M. Khapra, Sapan Shah. Projecting parameters for multilingual word sense disambiguation. 2009.

[24] Sudip Mittal, Prajit Kumar Das, Varish Mulwad, Anupam Joshi, and Tim Finin. Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pages 860–867. IEEE, 2016.

[25] Sudip Mittal, Anupam Joshi, and Tim Finin. Thinking, fast and slow: Combining vector spaces and knowledge graphs. *corpus*, 2:3, 2017.

[26] Mu Li Ming Zhou Nan Yang, Shujie Liu and Nenghai Yu. Word alignment modeling with context dependent deep neural network. 2002.

[27] NIST. Cybersecurity — nist. https://www.nist.gov/topics/cybersecurity/, 2018.

[28] The Global WordNet Organization. Wordnets in the world. http://globalwordnet.org/wordnets-in-the-world/.

[29] a Peter F. Brown, Jennifer C. Lai and Robert L. Mercer. Aligning sentences in parallel corproa. 1990.

[30] PyPI. Wiki-ru-wordnet. https://wiki-ru-wordnet.readthedocs.io/en/latest/, Sept 2017.

[31] 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.

[32] The Register. Make america late again: Us 'lags' china in it security bug reporting. https://www.theregister.co.uk/2017/10/20/us_china_vuln_reporting/, Oct 2017.

[33] Alexandra Birch Rico Sennrich, Barry Haddow. Neural machine translation of rare words with subword units. 2016.

[34] Deepali Vora Suraj Subramanian. Unsupervised text classification and search using word embeddings on a self-organizing map. 2016.

[35] Zareen Syed, Ankur Padia, M. Lisa Mathews, Tim Finin, and Anupam Joshi. UCO: A unified cybersecurity ontology. In *AAAI Workshop on Artificial Intelligence for Cyber Security*, pages 14–21. AAAI Press, 2015.