Semantically Rich Framework to Automate Cyber Insurance Services

Ketki Sane, Karuna Pande Joshi, *Senior Member, IEEE,* and Sudip Mittal, *Member, IEEE*

Abstract—With the rapid enhancements in technology and the adoption of web services, there has been a significant increase in cyber threats faced by organizations in cyberspace. It has become essential to get financial cover to mitigate the expenses due to a security incident. Organizations want to purchase adequate cyber insurance to safeguard against the third-party services they use. However, cyber insurance policies describe their coverages and exclusions using legal jargon that can be difficult to comprehend. Parsing these policy documents and extracting the rules embedded in them is currently a very manual time-consuming process. We have developed a novel framework that automatically extracts the coverage and exclusion key terms and rules embedded in a cyber policy. We have built our framework using Information Retrieval and Artificial Intelligence techniques, specifically Semantic Web and Modal Logic. We have also developed a web interface where users can find the best matching cyber insurance policy based on particular coverage criteria. To validate our approach, we used industry standards proposed by the Federal Trade Commission document (FTC) and have applied it against publicly available policies of seven insurance providers. Our system will allow cyber insurance seekers to explore various policy documents and compare the paradigms mentioned in those documents while selecting the best relevant policy documents.

Keywords Cybersecurity, Artificial Intelligence, Cyber Insurance, Knowledge Graph, Ontology, Knowledge Representation, Policies.

1 Introduction

There has been an increase in the number of cybersecurity threats that can lead to losses in-terms of sensitive PII, competitive data, and the potential reputation of an organization. An organization that is inflicted by a cyber-attack has to deal with costly expenses to mitigate the negative reciprocation. These costs generally include expenses related to technology upgrades, repairing a company's reputation, legal fees, etc. This results in the organizations move towards insurance providers to minimize some of these costs of potentially devastating effects after a cyber-attack.

Cyber insurance policies provide various coverages in case of a security breach that helps the organizations to reestablish the reputation loss and minimize various expenses. In the case of a security breach, the insurance provider is expected to pay the insurance amount to the organization. There have been innumerable cases in the past where the insurance providers decline to pay for a loss which they view to be exempted from the policy's security cover. These decisions are often contested by insurance providers and insured entities based on their intricate interpretation of the legal language stated in the policy.

In a situation where the insurance service provider refused to pay the insurance coverage amount due to disagreement between the two stakeholders, they often undergo courtroom discussions to validate the coverage clauses. Such scenarios showcase various loopholes and gaps present in the insurance policies. This mismatch of understanding between insurance providers and insured entities inspires little confidence in businesses while selecting the right cyber insurance vendor that could deliver on all the expected security coverages for their organization. The manual process of going through each and every vendor offering is time-consuming, as the process of analyzing every policy in detail is laborious. Also, every policy has a different definition and linguistic for its coverage clauses, so it is difficult to compare policy coverages from various policies. There is currently no good technical solution that can help to make this comparison easier. We have identified a need for a system that will be able to represent the contents of different cyber policies in a unifying common representation that can be automated for easy machine consumption. A complete understanding of legal nuances is important to perceive the elements of a cyber insurance policy. Also, there is a need for a system that can convert the insurance policy document to a machine-processable graph database which will reduce human efforts. We aim to build a system that would be able to suggest the best policy from a vendor that satisfies consumer's expected risk cover, coverage limits, and expected rate of coverage.

To solve this problem, we have developed an *ontological framework* that automates the population of digital policy documents into semantically rich knowledge representation. Our framework automates the process of extracting the policy key terms and their definitions and represents them in a format that the user can query upon. This model will help insurance seekers to find the best matching policy as per their requirements. Keyword searches may also return a huge number of expected matches, but that requires a lot of analysis and sorting of responses. The policy structure makes it difficult to clearly understand what is covered and what that coverage means. Multiple policy documents may

- *Ketki Sane and Karuna Pande Joshi are with the Department of Information System, University of Maryland Baltimore County. E-mail: {ketkij1, karuna.joshi}@umbc.edu*
- *Sudip Mittal is with Department of Computer Science and Engineering, Mississippi State University. E-mail: mittal@cse.msstate.edu*

list the same coverage but may not mean the same in terms of liability. Most of the policy documents are lengthy and dealing with these documents and understanding the key terms used in the document is a tedious task. To compare each policy coverage with another, one has to have complete knowledge about the domain and should be able to invest a lot of time matching the coverages and exclusions. To overcome these problems, building an ontological model for legal documents seemed to be the most efficient solution. This can help to capture legal key terms and rules in order to perform analytically and answer queries.

Our framework captures knowledge in form of key terms, rules, key terms descriptions, relationships between various legal terms, semantically similar terminologies, and deontic expressions. We created a semantically rich policy-based knowledge graph for Cyber Insurance using standards proposed by the United States Federal Trade Commission (FTC). We used Semantic Web technologies like OWL [18], RDF [17] and SPARQL [29], along with Natural Language Processing (NLP) and text mining techniques to extract the key features and query the knowledge graph. In this paper, we illustrate the knowledge graph in detail along with the main classes and relationships. In our research scope, we mainly focused on insurance policies coverages and exclusions. Based on the extracted key terms, we populated the knowledge graph which is a subset of the original knowledge graph consisting only inclusions and exclusions. We have validated this Knowledge Graph against the FTC (Federal Trade Commission) document [8] and publicly available insurance policies from seven insurance providers including [2] [3] [24] [25] [26] [27] and [28]. Our ontological model would also help the insurance providers in designing standardized policy documents.

We had published a paper earlier where we discussed about initial idea on building a system for automatically extracting coverages and exclusions from cyber policy documents, replacing the need of a human to manually interpret lengthy policy documents. In both the papers we use United States Federal Trade Commission (FTC) as a source of truth for comparing our results. In our previous paper, we discovered basic co-relations between the stakeholders involved in the cyber insurance domain. We worked on extracting all the coverages listed in the policy documents. We also discussed about building an interface for user to visualize and compare the policies and potentially employ a negotiation engine for procuring the best cyber insurance policy amongst different providers on behalf of the user.

This work is an extension of the work we did earlier. In the current work, we focus on extraction of exclusions, validation of extracted key terms and building a user interface to allow user to query the ontology knowledge graph. In this paper we focus on implementing a system for querying the knowledge graph using python's SparqlWrapper and Flask libraries. We also built a web interface using Angular, where a user can find the best matching policy by comparing the coverages and exclusions offered by each policy. Users can also see the description of how the selected coverage is defined in a particular policy documents. Fig. 6, Fig. 7 and Fig. 8 describes this application well.

We used Angular as our front end technology to design the user interface that runs a node-based JavaScript server (See Fig. 6). We used Bootstrap and JQuery components to style the user interface and to add dynamic behavior based on user interaction. To find the matching policy by querying the populated knowledge graph, we used python's SparqlWrapper and Flask libraries. Flask library spins up a lightweight web server where our microservice is hosted. Our microservice formulates the sparql queries using SparqlWrapper. The microservice then hits an endpoint at Fuseki server which has all the RDF triple assertions generated from the previous ontology service module. The Fuseki server executes all the SPARQL queries received from microservice POST HTTP calls and computes the result. This result is then passed on to the microservice which is inturn passed back to Angular applications ajax callback and displayed on the user interface.

The rest of the paper is organized as follows – In Section II, we discuss the related work in this area. Section III describes our framework for building and populating the knowledge graph. Section IV includes the results of our validation. We conclude in Section V.

2 Related Work

With the increase in magnitude of cyber security incidents across the spectrum, cyberspace has seen an increase in the number of players in the market extending insurance cover to online business firms for protecting their business. Romanosky et al. [8] have worked on studying cyber insurance providers in the market to analyze what processes the insurance providers have in place when they do security risk assessment for their clients and also what yard sticks, they use for formulating the pricing model. Their work attempts to throw light on maturity and effectiveness of models developed and followed by cyber insurance providers while formulating policies and cost structure for the clients. They have mainly focused on three key aspects: Firstly, representation of the risks covered by the provider and also identify exemptions made in the policy, Secondly, study of information sought by insurance providers to understand security standing of their client and thirdly, understand methodology for computing premiums. They concluded that different insurance providers had similar ideas of the risks that will be covered by their policy, but they saw significant variation in the coverage exemptions for each vendor's policy. Also, the study reveals that there are gaps in questionnaires posed by insurers which do not necessarily allow them to identify security standing of the third-party providers of their insured entity. Lastly, the study found lack of standardization on pricing models adopted by insurance providers. A very few policies seemed to take into account the security readiness of their insured entity, but largely the study found absence of any good heuristic model for price computation. Various stakeholders involved in the cyber insurance such as insurers and insured organizations need to exchange the important information such as queries or requests with the guarantee of sharing a common meaning. This can be achieved by using Semantic Web Technologies to model and reason about the services related information. We used Web Ontology Language (OWL) [18] and Resource Description Framework [17] to capture properties and relationships between the stakeholders and key elements in

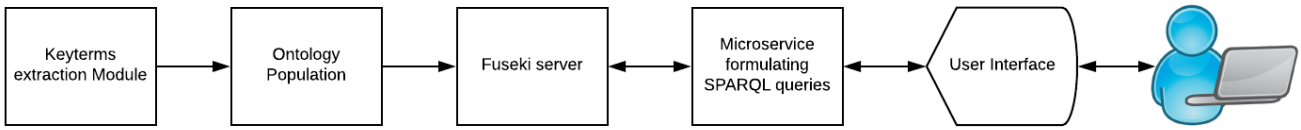

Fig. 1. Framework architecture detailing our knowledge extraction process, the extracted data is populated in our ontology hosted on a Fuseki server. The system also provides a user interface for queries.

cyber insurance policies. These semantic technologies help us to perform reasoning over our ontology.

Romanosky et al. [8] concluded that information asymmetry between insurer and insured is the primary reason for costly insurance premiums. In researcher's opinion, the cyber insurance policies fail to consider secondary losses in policy formulation, resulting in sub-optimal behavior of claims expected from their insured entities. Their work highlights the mindset of IT managers when they take decisions in reporting claims to their cyber insurance providers in lieu of potential secondary losses that the firm might incur in addition to the primary loss already incurred. On one hand, under reporting of cyber incident case details prevents the cyber insurance market from maturing and on the other hand, information asymmetry from the demand side leads to under utilization of policy. Thus, this work again calls outs lack of standardization in formulation of policies. This lack of common semantics between insurance providers and insured entities is the root cause of improper assessment of security posture of insured entities. This ripples down into incorrect coverage offerings. Also, the lack of standardization leads to non-sophisticated pricing models. Our work attempts to use semantic web techniques for modelling cyber insurance space. This approach allows one to develop fine grain models with ability to effectively reason over them.

Previous work done in [32], [33] and [34], have successfully demonstrated use of semantic web techniques in automating analysis and compliance of cloud service agreements, big data cloud policies and legal documents. These researches made use of GATE (General Architecture for Text Engineering) based approach leveraging different text mining and text extraction techniques. These systems similar to our work make use of domain specific ontologies to represent their space of work.

We have used Semantic Web to capture the properties and relationships between the stakeholders involved in the cyber insurance space. These semantic web technologies like Web Ontology Language (OWL) [18] and Resource Description Framework (RDF) [17] helped us asserting the knowledge from insurance domain and representing it with the domain-specific properties in a knowledge graph. We have performed reasoning over this semantic representation using SPARQL [29].

Bohme et al. [19] proposed a comprehensive unifying framework to illustrate the parameters that should be included in the modeling of cyber insurance. They focused on capturing relationships between information asymmetry, interdependent security, and correlated risk. They developed a framework to model all existing literature in cyber insurance space using a unifying model of terminology, that can capture all properties of the individual existing models. This approach resulted in a more robust framework that helped discover and alleviate shortcomings of previous models and their outcomes.

Ganino et al. [15] discussed about how plethora of information available via several open source data sources can be harnessed via automation of open Source intelligence using ontological models. In this work they have talked about the underlying usage of domain rich ontologies as being a key ingredient in their system architecture. Using ontology helped them to extract important key terms from unstructured public data sources available in digital form. Their work demonstrated successful use of ontology population pipeline including techniques developed for mining relevant information. Their validation framework also substantiated qualitative correctness of populated ontology. One of the important highlights of this research is that, 1) it showed how semantic web technologies can be used to reduce labor intensive manual tasks by developing a framework providing ability to extract structured information from structure free documents using ontology 2) And it also demonstrated use of semantic web languages, like SPARQL [29], to provide ability to query populated knowledge graph. Our work is also inspired by a similar technique where we have modeled cyber insurance policies using ontology and we have provided a web-based user interface application that harnesses semantic web tools and technologies like SPARQL [29] and RDF [17] to reason over our populated knowledge base.

3 Technical Approach for the Framework

In this section, we describe our approach towards developing a framework for automatic knowledge extraction and population of cyber insurance coverage and exclusion terms in the knowledge graph. Fig. 1, illustrates the overall architecture of our system. The architecture consists of 5 modules. The Keyterm extraction module uses Sentence Tokenizer and POS tagger for preprocessing of text. The Part of Speech tagger maps words in a sentence to their part of speeches. This way we get all words in the document with their respective part of speeches. Our deontic logic parser extracts key terms which we then compare with our knowledge graph from FTC [8] and generate a resulting knowledge graph containing all the coverage and exclusion keywords from policy documents. This knowledge graph is then represented in form of ontology. We use Fuseki server

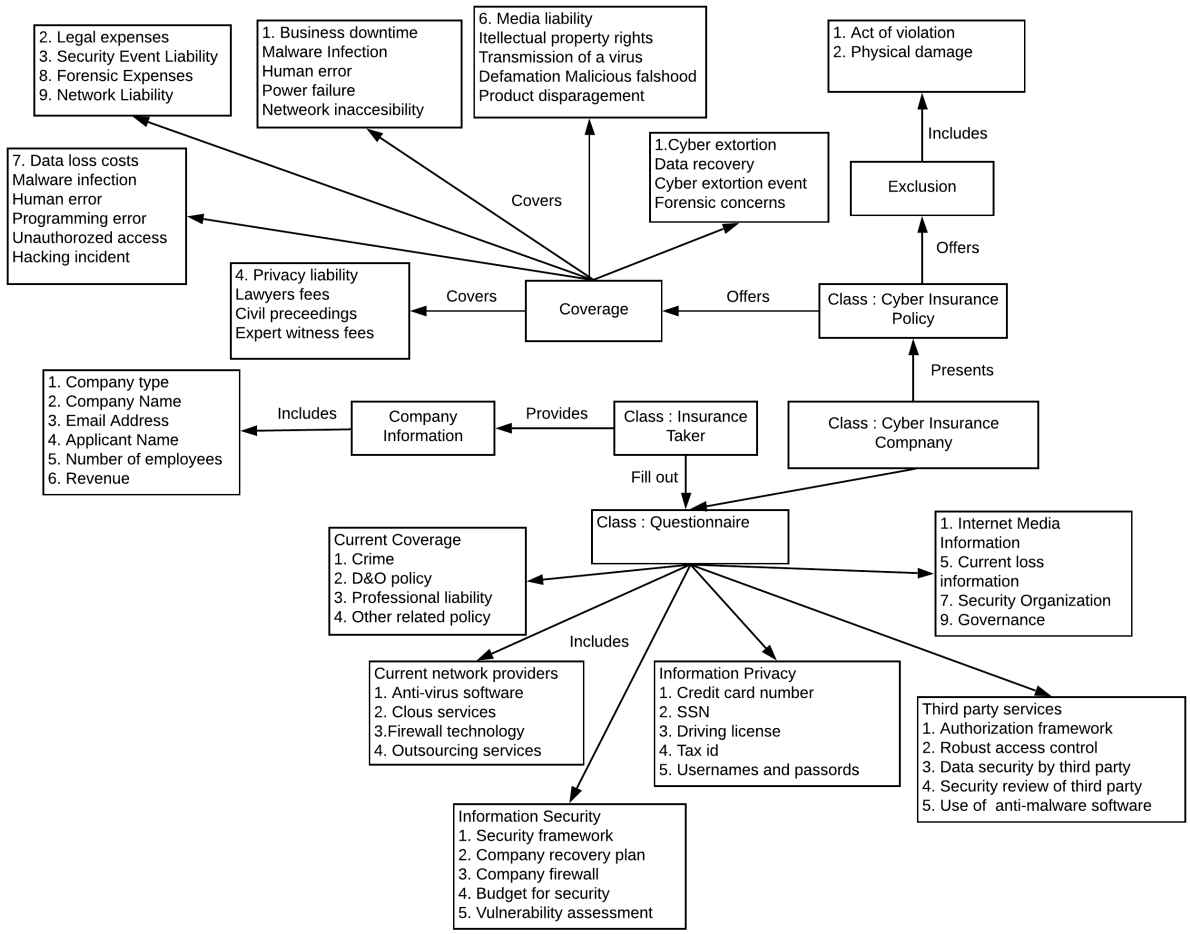

Fig. 2. Our knowledge graph represents key components in the cyber insurance policy and security questionnaire.

to store the ontology in form of RDF triples. It is then used to execute SPARQL [29] queries and compute the result. This result is displayed on the user interface built using Angular. On this web interface a user can see and compare multiple cyber insurance policies. This is described in detail in below sections.

We have build this framework using techniques from Natural Language Processing (NLP), Information Retrieval and Artificial Intelligence, specifically Semantic Web technologies. For our study, we collected cyber insurance policies, available in the public domain, of various insurance providers. For validation and ground truth, we used the analysis done by Romanosky et al. at the Federal Trade Commission [8]. Our system automatically extracts various coverages and exclusions from policy documents and inserts them in the cyber insurance knowledge graph.

Our framework consists of three key parts listed below and are described in detail in the following sections:

- **Knowledge graph for Cyber Insurance Service:** We studied the cyber insurance analysis report by the Federal Trade Commission (FTC) [8] as the ground truth for the cyber insurance domain and identified the main entities and topics described in the report. This helped us determine the key classes of our knowledge graph along with their relationships. Our knowledge graph, defines the relationship between the four main stakeholders in cyber insurance, viz. Cyber Insurance company, Cyber Insurance policy, Insurance seeker and Insurance company questionnaire.

- **Automated Text Extractor for Coverages and Exclusions:** This module automatically extracts key terms and rules describing coverages and exclusions from a policy document and populating them in form of RDF triples. We have divided our implementation for this module into two main components, namely, a text coverage/exclusion extraction module to extract ontology coverage and exclusion classes from cyber insurances policy documents, and a core service module to represent the knowledge graph in form of triples. Using deontic logic, we identified various coverages and exclusions. We applied the sentence tokenizer on a given policy document, and then further categorized the tokenized sentences as either a coverage or an exclusion. In order to find answers to questions like, 'What coverages this policy provides?',

'What coverages the policy won't pay for?' we further classify rule sentences into Permissions and Prohibitions.

- **Querying & Reasoning over the knowledge graph:** We have built a knowledge graph, using OWL, to represent the terms and rules embedded in the cyber insurance service policies. For hosting our knowledge graph as a service endpoint we used the Fuseki Web Server. We have also designed a User Interface that allows user to access our system and query the knowledge graph using semantic technologies languages like SPARQL [29].

3.1 Knowledge graph for cyber insurance

Our ontological model can help insurance provider organizations to create a structured machine processable cyber insurance policy rules which can be automatically parsed and queried upon. We studied 7 most popular insurance policies. The policies include [2] [3] [24] [25] [26] [27] and [28]. We referred to the Federal trade commission (FTC) document [8] to identify key classes of our cyber insurance ontology along with their relations. We developed an ontology that can capture the insurance policy key components and illustrates their definitions, associated rules and types. Our knowledge graph, defines the relationship between the four main stakeholders, Cyber Insurance company, Cyber Insurance policy, Insurance seeker and Insurance company questionnaire in a cyber insurance environment. The detailed knowledge graph is illustrated in Fig. 2. We designed this knowledge graph using Protege software [23] and populated it by using Apache Jena [31].

Based on the study of insurance products we have our main 4 classes as 'Cyber Insurance Company', 'Insured', 'Cyber Insurance Company Questionnaire', and 'Cyber Insurance Policy' respectively. The classes are described in detail as follows:

1) **Cyber Insurance company**: This class describes the insurance vendor company that offers the cyber insurance policy.
2) **Cyber Insurance policy**: This class represents the written policy document. Every insurance policy has rules for coverages and exclusions. The Cyber insurance policy class has 4 main subclasses viz. Coverages, Exclusions, Cyber Insurance Questionnaire, and Insured / Cyber Insurance seeker. The inclusions or coverages are the key aspects that the policy provides. Each policy document has coverages and exclusions that represent what the policy covers and what they do not cover.
3) **Cyber Insurance Questionnaire**: The questionnaire class captures all the details that the insurance seeker must fill out while opting for an insurance policy. The questionnaire has set of questions such as details on the applicant's current security infrastructure, details about any third-party services used by the applicant, security frameworks in use, etc. All the information submitted by Insurance seeker is used by the insurer to decide on the price, appropriate coverage elements for that Insurance seeker.
4) **Insured / Cyber Insurance seeker**: The applicant who is willing to opt for the insurance is categorized as Insurance seeker or Insured. Insured provides all the information required by the insurer. Usually, the insurance seeker is asked to provide specific details such as the organizations name, address, number of employees, number of third-party vendors, etc.

Following relationship were also modeled in the ontology:

- *offers*: Cyber Insurance Company → Cyber Insurance Policy.
- *presents*: Cyber Insurance Company → Cyber Insurance Company Questionnaire.
- *has*: Cyber Insurance Policy → Coverages, Exclusions.

In our study, we observed that there is no standard format for cyber insurance documents. There are many organizations providing cyber insurance and each insurance provider structures its policies in its own format. By capturing the key components of cyber insurance policies as a knowledge graph, we can facilitate automatic comparison of two or more policy offerings, thereby enabling consumers to make calculated choices.

3.2 Automated Text Extractor for coverages

The cyber insurance policies are very complex in nature. It is very time consuming to manually go through all the documents and find the listed coverages and exclusions. There is no standard established in the way policy should be structured. In case of security incident, the complexity and ambiguity of policy documents makes it difficult to understand what is covered and this might make the stockholders to resolve their discrepancy in the court. We believe that, our ontological approach for representing the insurance policies, would help reduce many of these issues. Having a framework that automates the mechanism of extraction of key elements from a policy document is beneficial. We have categorized extracted rules on coverages (permissions) and exclusions (prohibitions) using deontic expressions. Following is the grammatical regular expression [1] used for permissions / prohibitions:

- $<Pronoun\ |Delimiter><deontic><Noun\ phrase>$

Permission	Prohibition
Will incur	Exclude
Will cover	Not provide
Be liable	Not liable
Will pay	Not include

TABLE 1
Modal verbs in Deontic Expressions.

Table 1, lists the modal verbs we considered when classifying rules into Permissions and Prohibitions. Modal logic [41] is a broad term used to cover various other

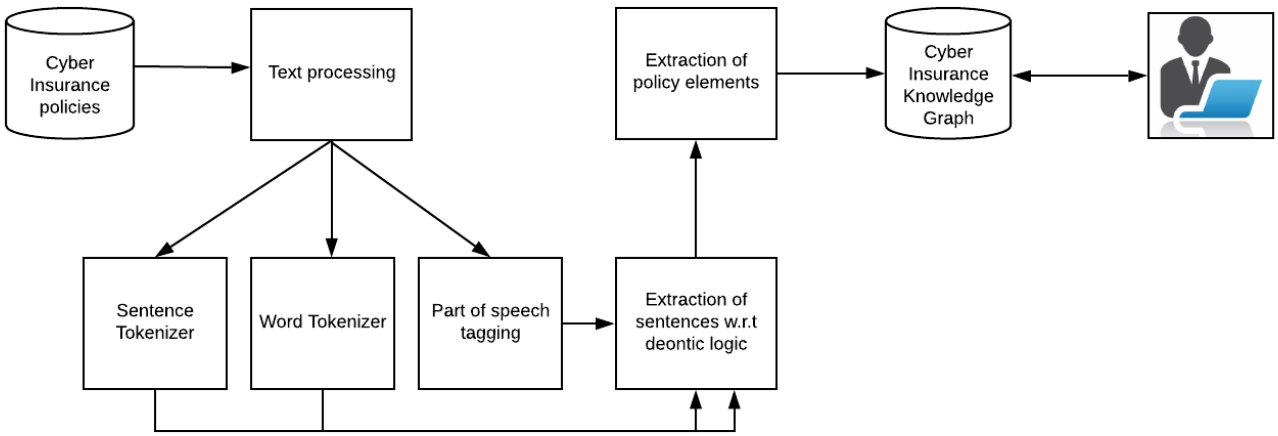

Fig. 3. System architecture of coverage extraction automation system. The system takes as an input various cyber insurance documents and processes them to extract policy elements.

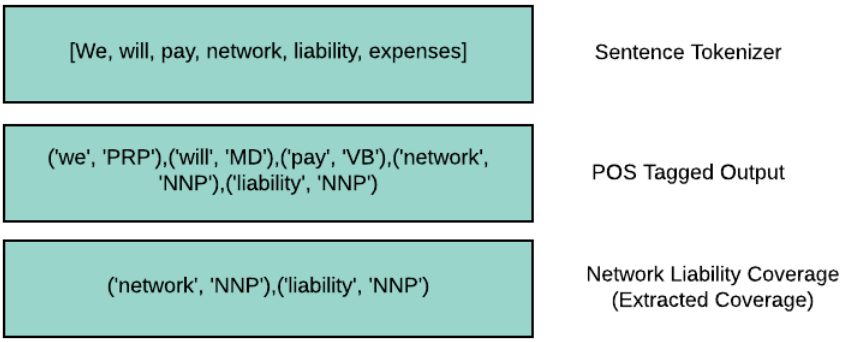

Fig. 4. Example showcasing preprocessing of text using sentence tokenizer and a part-of-speech tagger.

forms of logic such as temporal logic and deontic logic. Deontic logic describes statements containing permissions, and obligations, and temporal logic describes time-based requirements. Deontic logic further consists of four types of modalities:

- *Permissions* (included coverages)
- *Prohibitions* (stated exemptions)
- *Obligations* (Mandatory conditions for extending a coverage)
- *Dispensation* (Non-mandatory conditions for extending coverage)

We used deontic Logic to further identify various coverages and exclusions. After tokenizing sentences in a given policy document, we categorized them as either a policy coverage or an exclusion and populated the coverage/exclusion classes. In order to find answers to questions like, *What coverages does this policy provide?*, we need to classify sentences into the two categories.

Our coverage extraction module, illustrated in Fig. 3, uses a grammar chunking parser, which takes deontic grammar rules as input for permission and prohibition. Given a policy document, using the Natural Language Toolkit (NLTK) sentence tokenizer [39], the system gets unique sentences in a policy document. Fig.4, shows some of the outputs generated.

We then use a word tokenizer on each of the sentences to get a list of words in the sentence. The output of the tokenizer is served as the input to POS (Stanford part of speech tagger). The POS tagger tags each word in the sentence with respect to the part of speech. For example, the POS tagged output for a sentence in the policy such as, "We will pay Third party liability expenses" will be: "('we', 'PRP'), ('will', 'MD'),('cover, 'VB')('Cyber, 'JJ'), ('extortion, 'NN'),('liability', 'NN') where PRP is a preposition, MD is a modal, VB is a verb, JJ is an adjective, and NN is a noun.

Earlier, for extracting coverages and exclusions from the policy document, we tried a text based regular expression (RegEx) [13] parser to find all the sentences in a policy that pertain to coverages and exclusions. This approach did not work well, as regular text RegEx conforms to a narrow set of textual patterns and is less flexible. We found that using a grammar-based natural language chunking parser is better suited to solve this kind of problem. We next explored use of deontic expressions on the policy documents to extract policy coverages/exclusions.

3.3 Automated Text Extractor for Exclusions

In this section, we describe our approach for extracting exclusions from cyber insurance policies. For extracting coverages, we used various NLP techniques as described in earlier section. For extracting exclusion, we tried the similar approach of deontic logic at first, but, it did not work well for policy exclusions. In all of the seven policy documents under consideration we found that, none of the policy document conformed to a particular deontic logic or grammar pattern in the policy wording. For example, in the policy document named as Axis Capital [28], the policy wording to state that there will be no coverage provided in case of Bodily Injury, is as follows: "Bodily Injury except that this exclusion does not apply to mental injury or mental anguish if directly resulting from an Enterprise Security Event involving Protected Personal Information that gives rise to an Enterprise Security Event Claim". While in another policy named working of XL-Catlin [26], they have illustrated the same exclusion as "Bodily injury, sickness, disease, emotional distress, mental injury, mental tension, mental anguish, pain and suffering, humiliation or shock sustained by any person, including death that results from any of these, or damage to or destruction of any tangible property, including loss of use thereof whether or not it is damaged or destroyed; provided, however, this exclusion will not apply to any otherwise covered claim for emotional distress, mental injury, mental tension or mental anguish, pain and suffering, humiliation or shock that directly results from a covered third party wrongful act." We found it technically difficult to map both of the sentences to one deontic expression. Also, there is no particular grammatical pattern followed by these policy documents, so it was difficult to parse the policy document through any type of grammatical regex.

Our aim was to locate the exclusion keywords from policy document, extract them and represent them in a particular format which is query-able. The next approach that we tried was to locate paragraphs in cyber insurance policy that talk about exclusions. In this way, after locating the exact exclusion paragraphs, we planned on applying some text extraction techniques to find the exclusion keywords After researching though the best techniques for paragraph extraction in a digital document, we tried a few python libraries like 'gensim', 're'. Using these libraries, we first attempted to find the Start and Stop patterns for exclusion paragraphs for each policy documents. Unfortunately, this approach failed because although every document has similar kind of start pattern, but the stop pattern is different for each of them. For example, in all seven policy documents, the exclusion sections were titled as follows: 'General Exclusions' or 'Exclusions' or 'What is not covered'. For finding stop pattern for exclusion section, we thought of two possible options. First, to find the last sentence of exclusion paragraph, second, to find the title or first sentence of next section in the policy document. But, as every exclusion section ends with a non-definite end pattern and also as the order of sections in a cyber policy document is not deterministic, we failed to find a robust solution to find stop pattern for exclusion section. Due to this bottleneck, we did not pursue this technique ahead.

Next, we explored use of n-grams model for exclusion extraction. We identified the exhaustive list of exclusions form the FTC document [8] and also from the seven policy documents. Then, we ran our policy documents to though our ngram extractor module to extract bi-grams, trigrams and four-grams from the policy documents. We did some post-processing of ngram tuples to map the ngram with our exhaustive list of exclusions. Fig. 5 displays our automated exclusion extraction architecture.

3.4 Population of ontology

Once we get coverage sentences using grammatical regEx or exclusion key terms using Ngrams, the next step is to map the coverage or exclusion to appropriate ontology classes.

Using deontic grammar parser, we got the sentence partitioned a *subject*, *predicate*, and *object*. To map the extracted coverage sentence to particular ontology class, we used bag of words approach. For example, if a policy document has a coverage clause worded as, "The policy covers cyber extortion damages" and a similar clause in another policy worded as, "We will pay the costs you incur subject to cyber breach", this sentence will be mapped to a coverage class called "Cyber Extortion". Each policy document can refer the same coverage with different name. For example, Chubb policy [3] refers Cyber related expense as 'Cyber Extortion' while Hiscox policy [2] refers the same coverage as 'Cyber Business Interruption'. To capture all these details, we created a static mapping of different key terms to a single standard term defined by FTC [8].

Table 2, shows most common coverage terms mapped to its respective coverage term. Our python-based extractor module compiles all the coverages and exclusions identified in a policy into a json payload which is sent further to our ontology service. The ontology service (Ontology population module) uses an open source semantic web framework for Java (JenaApi) [31] and creates class individuals. Once the all the individuals are created successfully, it is ready to accept a user query. Our ontology service stores serialized Resource Description Framework (RDF) [17] triples in-memory.

3.5 Querying & Reasoning over the knowledge graph

We have developed a web-based cyber insurance application for providing a user interface to specify a set of requirements. Based on user requirements, the system will retrieve the best matching cyber insurance service. Using this interface, users can select the service they are interested in and drill down to see the coverages/ exclusions listed in that policy. Fig. 8 illustrates the coverages and exclusions listed in Chubb [3].

Fig. 8 shows the query interface for our system that compares multiple coverages with respect to their policies. User can select whatever coverages he wants to have in his policy, and the system lists out the policies that cover these coverages. These policies states the sentences that explains the coverage terms in detail. As shown in Fig. 6, user can select the coverages of his choice as Network Liability, Media liability and so on. After selecting these coverages, he can find the best matching policies as per his requirements. Fig. 7 shows the coverages provide by

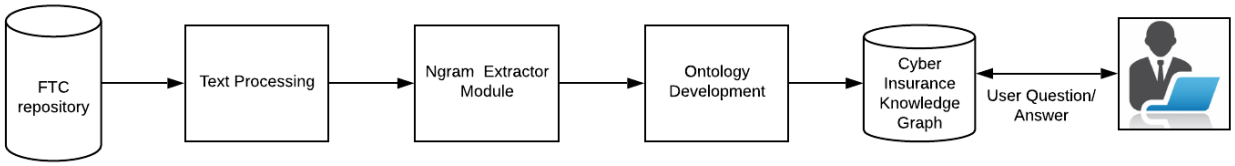

Fig. 5. System Architecture of exclusion extraction automation system.

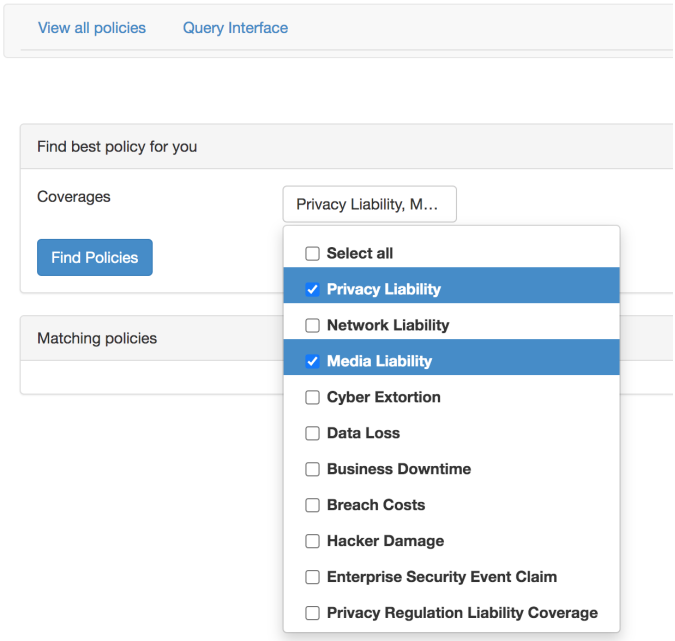

Fig. 6. The system also includes a selection of coverages allowing the user to create complex queries.

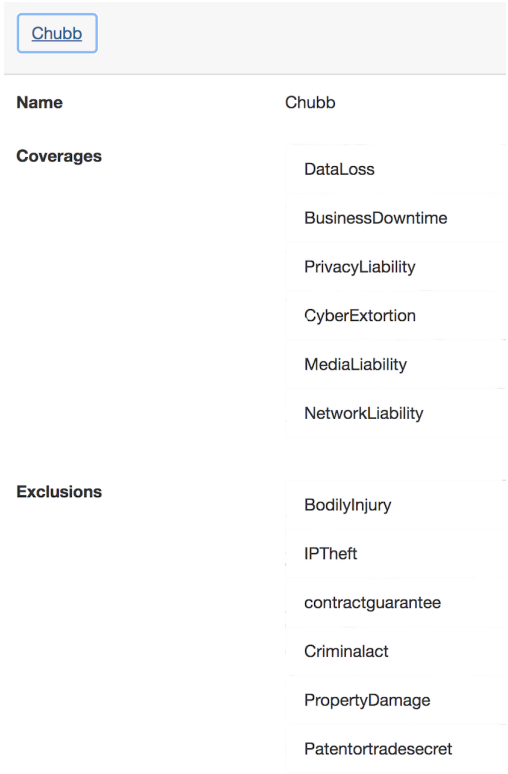

Fig. 8. Extracted coverages and exclusions in the Chubb Insurance Policy.

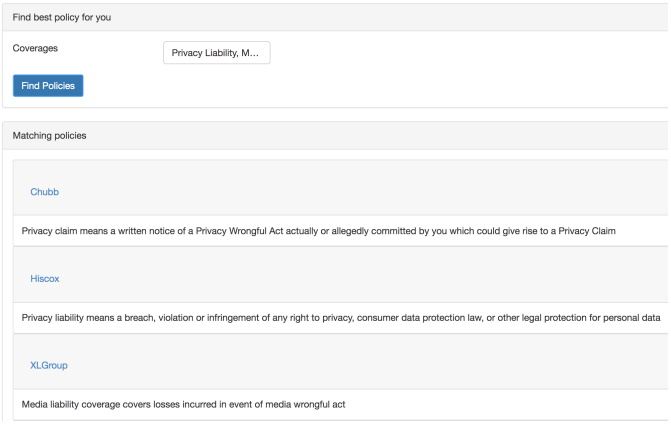

Fig. 7. Query interface allows a user to create complex SPARQL queries.

the policies with description. For example, in Fig. 7, if user selects Privacy liability and Media liability as coverages, the policies that cover them will show what these terms mean. There are three policies Chubb, Hiscox and XL Group that provide Privacy liability and Media liability as their coverages. The Privacy liability coverage is explained Chubb policy as 'Privacy claim means a written notice of a Privacy Wrongful Act actually or allegedly committed by you which could give rise to a Privacy Claim' whereas the same coverage is stated as 'Privacy liability means a breach, violation or infringement of any right to privacy, consumer data protection law, or other legal protection for personal data' in Hiscox policy. The XL Group provides coverage in terms of Media liability and states it as 'Media liability coverage covers losses incurred in event of media wrongful act'.

We used Angular as our front end technology to design the user interface that runs a node-based JavaScript server (See Fig. 6). We used Bootstrap and JQuery components to style the user interface and to add dynamic behavior

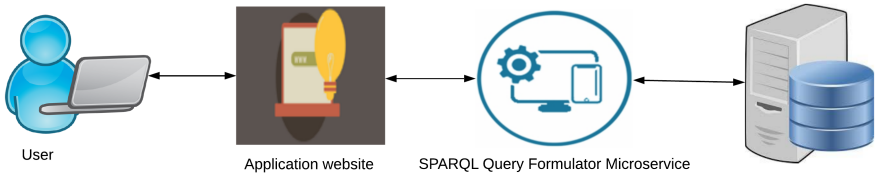

Fig. 9. User interface architecture enables complex query input through a website, which in turn executes a SPARQL query microservice. The microservice interacts with our Fuseki server.

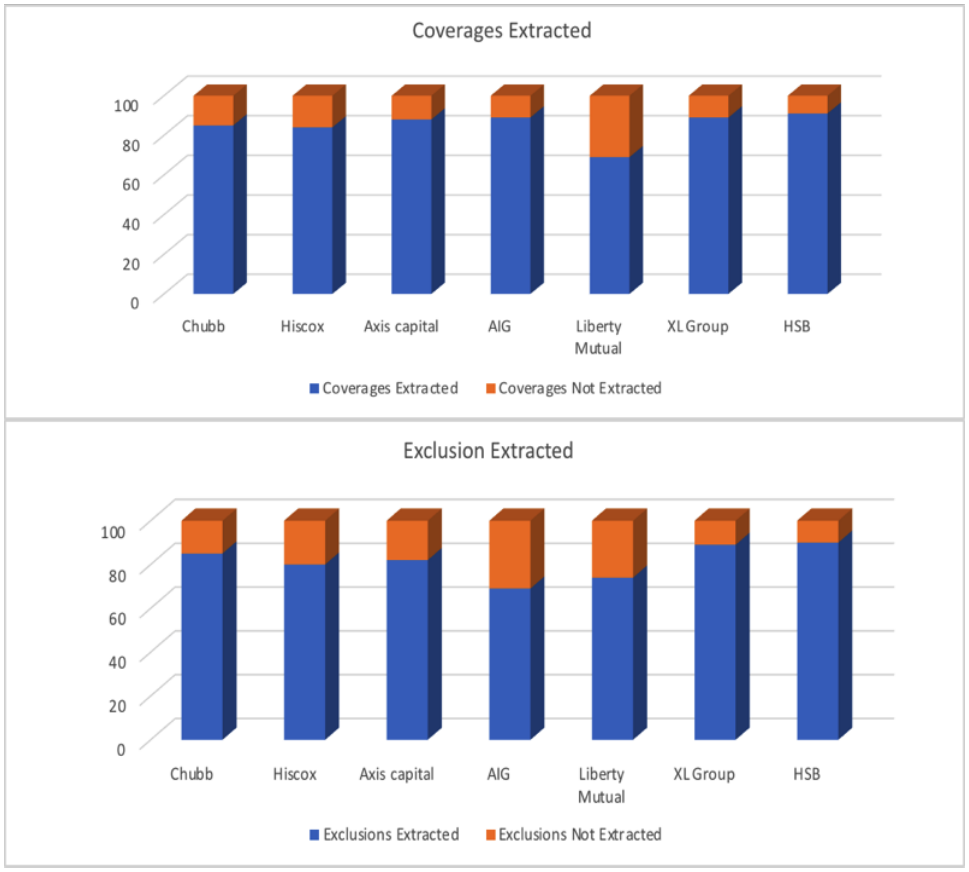

Fig. 10. Bar chart showing percentage of extracted coverages and exclusions in different cyber insurance policies.

based on user interaction. To find the matching policy by querying the populated knowledge graph, we used python's SparqlWrapper and Flask libraries. Flask library spins up a lightweight web server where our microservice is hosted. Our microservice formulates the sparql queries using SparqlWrapper. The microservice then hits an endpoint at Fuseki server which has all the RDF triple assertions generated from the previous ontology service module. The Fuseki server executes all the SPARQL queries received from microservice POST HTTP calls and computes the result. This result is then passed on to the microservice which is inturn passed back to Angular applications ajax callback and displayed on the user interface.

4 Results and Validation

4.1 Extracted Coverage and Exclusion Key terms

We extracted the sentences that describe coverages and exclusions and satisfy the deontic expression linguistic structure for various permissions and prohibitions. Table 3, shows some of the most common coverages from policy documents with extracted sentences and mapping coverage classes for Chubb insurance policy. For example as shown in Table 3, coverage category for a sentence, 'We will pay Damages and Privacy Claims Expenses by reason of a Privacy Claim first made during the Policy Period', is 'Privacy Liability'. Similarly, we found all coverage sentences from available policy documents and identified the coverage category for them. We categorized sentences with the model verbs such as 'will incur', 'will pay', will cover' and be liable' as permissions while the sentences having model

Extracted Policy Coverages		
Policy Name	Coverages	Exclusions
Chubb	Privacy Liability, Network Liability, Media Liability, Cyber Extortion, Data Loss, Business Downtime	IP Theft, Criminal Act, Bodily Injury, Contract Guarantee, Property Damage, Patent or Trade secret
AIG	Breach of Personal Information, Breach of Corporate Information, Outsourcing, Data Administrative Investigation, Data Administrative Fines, Repair Company Reputation, Notification to Data subjects, Monitoring, Emotional Distress, Cyber Extortion	IP Theft, Criminal Act, Bodily Injury, Contract Guarantee, Contractual Liability, Property Damage
Axis Capital	Enterprise Security Event claim, Privacy Regulation Liability Coverage, Cyber extortion	IP Theft, Criminal Act, Bodily Injury, Contract Guarantee, Contractual Liability, Property Damage
Hiscox	Breach Cost, Hacker Damage, Cyber Extortion, Privacy Liability	Act of War, IP Theft, Criminal Act, Bodily Injury, Property Damage, Natural Disaster, Patent or Trade secret, Claims outside the applicable courts
HSB	Business Downtime, Extra Hardware, Recovert of Hardware, Waste Disposal cost, Attending courts, Removing Data, Hiring Professional Consultants, Investigation Costs, Loss Prevention Measures, Cyber Crime	Act of War, IP Theft, Criminal Act, Contract Guarantee, Natural Disaster
XL-Catlin	Technology And Professional Services, Media Liability, Privacy And Cyber Security, Privacy Regulation Liability Coverage, Supplemental Third Party Liability Prevention, Data Loss, Cyber Extortion, Business Downtime	IP Theft,Bodily Injury, Contract Guarantee, Non-monetary Relief, Fund Transfer, Property Damage, Seizure by government
Liberty Mutual	Cyber Extortion, Privacy Regulation Liability Coverage, Network Liability, Media Liability, Repair Company Reputation, Privacy Regulation Liability Regulation	IP Theft, Bodily Injury, Property Damage

TABLE 2
Extracted policy coverages and exclusions.

verbs such as 'not provide', 'not incur' were categorized as prohibitions. Table 4, shows all the coverages and exclusions extracted with the policy vendors. Table 1, shows modal verbs in deontic expressions. The output shows the extracted converges and exclusions for the policy documents

Extracted Policy Coverages for Chubb	
Coverage category	Extracted sentence
Privacy claim coverage	We will pay Damages and Privacy Claims Expenses by reason of a Privacy Claim first made during the Policy Period.
Network Incident coverage	We will be liable for Network Security Claims Expenses, by reason of a Network Security Claim first made during the Policy Period.
Media expense coverage	We will pay Media Claims Expenses and Damages by reason of a Media Claim first made during the Policy Period.
Cyber Extortion coverage	We will incur Cyber Extortion Damages in case of security breach.

TABLE 3
Extracted coverage sentences and classes.

[2] [3] [24] [25] [26] [27] and [28].

4.2 Ontology Population

We then populated our RDF triples in form of ontology which user can query upon. We populated individuals from all seven policies, and asserted all the objects and data properties corresponding to each individual. We used RDF triple assertions in this populated ontology and uploaded these assertions to Fuseki server. We populated the ontology with the classes that are subsets of the High-level ontology. We also injected the dependencies in terms of object properties and data properties such as

- policy → covers → coverageName
- Policy → hasDescriptionFor → Description

4.3 Validation

We validated the results of the populated ontology and extracted coverages and exclusions with following measures:

4.3.1 Evaluation Qualitative

Accuracy is a criterion that states if the definitions, descriptions of classes, relationships, properties, and populated individuals in an ontology are correct." [38] Completeness measures if the domain of interest is appropriately covered in the knowledge graph." [38] Adaptability measures how far the ontology and User Interface anticipates its uses. An ontology should offer the conceptual foundation for a range of anticipated tasks." [38] Clarity states how effectively the ontology communicates the intended meaning of the extracted legal terms." [38] Computational efficiency measures the ability of the used tools to work with the ontology, the speed that reasoners need to fulfill the required tasks." [38] Consistency describes if the ontology has any contradictions." [38]

4.3.2 Evaluation Quantitative

To check the accuracy rate, we used precision and recall methods. We checked the results relevancy with respect to

Coverage Extraction from Sentences		
Policy Name	Extracted sentence	Coverage
Chubb	We will pay Damages and Privacy Claims Expenses by reason of a Privacy Claim first made during the Policy Period	Privacy Liability
Chubb	We will pay Damages and Network Security Claims Expenses, by reason of a Network Security Claim first made during the Policy Period	Network Security Liability
Axis Capital	The Insurer will pay the Insured Entity for Extortion Loss incurred because of an Extortion Threat, in excess of the applicable retention and within the applicable Limits of Insurance	Computer System Extortion Coverage
Hiscox	we will pay all the reasonable and necessary expenses incurred with our prior written consent in replacing or repairing your computer system, programmers or data you hold electronically to the same standard and with the same contents before it was damaged, destroyed, altered, corrupted, copied, stolen or misused	Hacker damage
Hiscox	If during the period of insurance, and in the course of your business or advertising, you receive an illegal threat, we will pay the cost of any ransom demand from the third-party or, if the demand is for goods or services, their market value at the time of the surrender	Cyber Extortion

TABLE 4
Coverage Extraction from various policies.

correctly classified number of records and number of genuinely relevant results that are an outcome of this method. Fig. 9 shows the percenatge of extracted coverages and exclusions per policy. Based on the results obtained, we got the Precision, Recall and Accuracy for Coverages and Exclusions in percentage as follows: For coverages, results obtained for the above measures are: Accuracy = 65.12, Precision= 71.86 and Recall= 86.52. For Exclusion, we got Recall = 81.49, Precision = 66.00 and Average = 57.41.

5 Conclusion

Cyber insurance policy documents are hard to comprehend and to go through due to their ambiguous and complex nature. We have developed a semantically rich framework for automation of cyber policy documents. Our system automatically extracts the relevant deontic expressions and primary sentences explaining the policy expressions. In this paper, we have also developed a semantically rich ontology that captures information about cyber insurance providers, insurance seekers, key policy elements and questionnaire. The system provides a web user interface that servers as a platform for finding best policy based on dynamic coverage criteria provided by the user during run time. Automatic extraction and population of ontology from insurance policy documents along with querying platform can replace the manual work of analyzing the policy documents. This is the first step towards building a system that can interface with vendors and handle automatic extraction and storage of key terms of cyber insurance policy. Also, insurance providers can use this ontology as a blueprint to design and structure their policy offerings. This will effectively bring standardization in cyber policy documents across vendors.

In the future, this work can be extended to automate the workflow process of other three stakeholders involved in cyber insurance space i.e. insurance company, insurance policy and company questionnaire. In this study, we have developed a system where given required coverages, the system would give best matching policy. This can be extended by allowing user to select the inclusion limits and expected rate, and the system will find the best suitable policy.

6 Acknowledgment

This research was partially supported by a DoD supplement to the NSF award 1747724, Phase I IUCRC UMBC: Center for Accelerated Real time Analytics (CARTA).

References

[1] Karttunen, L., Chanod, J.-P., Grefenstette, G., and Schiller, A. (1996). Regular expressions for language engineering. Journal of Natural Language Engineering, 2(4):305–328.
[2] Hiscox PLC. (2015). Cyber and data Policy wording, WD-PIP-UK-CD(2) 13388 05/15.
[3] Chubb Insurance policy. https://www.chubb.com/cz-cz/assets/documents/chubbpp-cyber-enterprise-risk-management-en.pdf.
[4] Marthie Grobler, J.C. Jansen van Vuuren, Louise Leenen ImplementationofaCyberSecuritypolicyinSouthAfricaImplementationofaCyberSecuritypolicyinSouthAfrica.pdf
[5] James Geller, Soon Ae Chun, Arwa Wali, "A Hybrid Approach to Developing a Cyber Security Ontology", Proceedings of 3rd International Conference on Data Manage-ment Technologies and Applications, pp. 29-31, August (DATA 2014).
[6] N. Gcaza, R. V. Solms, and J. V. Vuuren. An Ontology for a National CyberSecurity Culture Environment. Proceedings of the Ninth International Symposium on Human Aspects of Information Security & Assurance (HAISA 2015), (Haisa):1– 10, 2015.
[7] Amina Souag, Camille Salinesi and Isabelle Comyn-Wattiau2 Ontologies for Security Requirements: A Literature Survey and Classification
[8] S. Romanosky, L. Ablon, A. Kuehn, T. Jones, Content analysis of cyber insurance policies: How do carriers write policies and price cyber risk?, 2017.
[9] Insurance Information Institute. (n.d.-a). Insurance Industry at a Glance. III. Retrieved from http://www.iii.org/fact-statistic/industry-overview
[10] Meenachi, N.: Web ontology language editors for semantic web-a survey. International Journal of Computer Applications 53(12), 12–16 (2012)
[11] Delia Rusu, Lorand Dali, Blaz Fortuna, Marko Grobelnik, and Dunja Mladenic. 2007. Triplet extraction from sentences. Proceedings of the 10th International Multiconference "Information Society – IS 2007", A:218–222, October.
[12] Stanford POS Tagger. http://nlp.stanford.edu/lexparser.shtml.
[13] Regular Expression Parser. https://dzone.com/articles/a-guide-to-parsing-algorithms-and-technology-part.
[14] Word and Sentence Tokenizer. https://training-course-material.com/training/Natural Language Processing in Python 5.2.Chinking.C2.B6.
[15] Ganino G, Lembo D, Mecella M, Scafoglieri F. Ontology population for open-source intelligence: A GATE-based solution. Softw Pract Exper. 2018;1–29. https://doi.org/10.1002/spe.2640

[16] Jansen van Vuuren J, Leenen L, Zaaiman J. Using an ontology as a model for the implementation of the national cybersecurity policy framework for South Africa

[17] O. Lassila, R. Swick, Resource Description Framework (RDF) Model and Syntax Specification, Feb. 1999.

[18] D.L. McGuinness, F. van Harmelen, "OWL Web Ontology Language Overview", Feb. 2004. Consortium, 2004.

[19] Bohme, R. and Schwartz, G. 2010. Modeling cyber-insurance: Towards a unifying framework. In Proceedings of the Workshop on the Economics of Information Security (WEIS).

[20] L. Ma, J. Mei, Y. Pan, K. Kulkarni, A. Fokoue, and A. Ranganathan, "Semantic web technologies and data management," in Proc. of W3C Workshop on RDF Access to Relational Databases, 2007.

[21] Oltramari, A., Cranor, L.F, Walls, R., McDaniel, P., "Building an Ontology of Cyber Security", in STIDS 2014 (9th International Conference on Semantic Technology for Intelligence, Defense, and Security, 2014

[22] Oltramari, Alessandro, Diane S. Henshel, Mariana Cains and Blaine Hoffman. "Towards a Human Factors Ontology for Cyber Security." STIDS (2015).

[23] Musen, M.A. The Protege project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI: 10.1145/2557001.25757003.

[24] AIG Insurance policy. http://www.aig.com/content/dam/aig/america-canada/us/documents/business/cyber/cyberedge-cyber-liability-insurance-brochure.pdf

[25] HSB Insurance policy. https://www.constructaquote.com/media/1519/hsbeil_cyber_policy_wording.pdf

[26] XL Catlin Insurance policy. https://xlcatlin.com/-/media/xlinsurance/pdfs/professional/cyber-liability/xl-catlin_ifl_apac_cyber-product-sheet.pdf

[27] Liberty Mutual Insurance policy. https://www.libertyspecialtymarkets.com/wp-content/uploads/2015/01/LSM074_Product_overview-Cyber_risksSCREEN.pdf

[28] Axis capital Insurance policy. https://www.axiscapital.com/docs/default-source/docs/insurance/us/professional-lines/

[29] Prud'hommeaux, E. and Seaborne, A. 2008. SPARQL query language for RDF. W3C recommendation. http://www.w3.org/TR/rdf-sparql-query/.

[30] OConnor, M., & Das, A. (2009). SQWRL: A query language for OWL. Proceedings of OWL: Experiences and Directions (OWLED), the fifth International Workshop.

[31] Apache Jena API. https://jena.apache.org/documentation/inference/

[32] Aditi Gupta, Sudip Mittal, Karuna P Joshi, Claudia Pearce and Anupam Joshi. Streamlining management of multiple cloud services. 2016 IEEE 9th International Conference on Cloud Computing (CLOUD).

[33] Karuna P Joshi, Aditi Gupta, Sudip Mittal, Claudia Pearce, Anupam Joshi and Tim Finin. Semantic approach to automating management of big data privacy policies. 2016 IEEE International Conference on Big Data (Big Data).

[34] Karuna P Joshi, Aditi Gupta, Sudip Mittal, Claudia Pearce, Tim Finin and others. ALDA: Cognitive assistant for legal document analytics. 2016 AAAI Fall Symposium Series.

[35] Bandyopadhyay, Tridib, Vijay S. Mookerjee, and Ram C. Bao. Why IT managers don't go for cyber-insurance products, Communications of the ACM - Scratch Programming for All 52, no. 11 (2009): 68-73.

[36] Equifax Breach. https://www.insurancejournal.com/news/national/2018/03/04/482301.htm

[37] NotPetya Effect. https://www.schneier.com/blog/archives/2019/02/cyberinsurance_.html

[38] Brank, J, Grobelnik, M, Mladenic, D. A survey of ontology evaluation techniques. In Proceedings of the conference on data mining and data warehouses (SiKDD 2005)

[39] NLTK https://github.com/nltk/nltk

[40] Fuseki Web server https://jena.apache.org/documentation/fuseki2/

[41] Modal Logic Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/logic-modal/

Ketki Sane graduated with a Master's degree in Information Systems from University of Maryland, Baltimore County in May 2020. She now works as a software engineer in IT at SemanticBits, LLC.

Karuna Pande Joshi is an Associate Professor in the Department of Information Systems at UMBC. She is the UMBC Site director for the Center of Accelerated Real Time Analytics (CARTA). She also directs the Knowledge, Analytics, Cognitive and Cloud Lab. Her primary research area is Data Science, Legal Text Analytics, Cloud Computing and Healthcare IT. She has published over 60 scholarly papers. Dr. Joshi has been awarded research grants by NSF, ONR, DoD, Cisco and GE Research. She was also awarded the NSF I-Corps award and TEDCO MII grant to explore commercial opportunities for her research and created a start-up on Data Science and Cloud technologies. She received her MS and Ph.D. in Computer Science from UMBC, where she was twice awarded the IBM Ph.D. Fellowship. She did her Bachelor of Engineering (Computers) from University of Mumbai. Dr. Joshi has also worked for over 15 years in the Industry primarily as an IT Project Manager. She worked as a Senior Information Management Officer at the International Monetary Fund for nearly a decade.

Sudip Mittal is an Assistant Professor in the Department of Computer Science & Engineering at the Mississippi State University. He received a Ph.D. in Computer Science from the University of Maryland Baltimore County in 2019. His primary research interests are cybersecurity and artificial intelligence. His goal is to develop the next generation of cyber defense systems that help protect various organizations and people. He is a member of the ACM and IEEE societies. His work has been funded by the NSF and Department of Defense.