University of Ibadan
Journal of Science and Logics in ICT Research

Development of an Extended Natural Language Processing (NLP)-based Framework For Knowledge Discovery In Terrorism-based Communication

Bolanle F. Oladejo **Noble I. Onyemenam**

Department of Computer Science, University of Ibadan, Ibadan, Nigeria.

Abstract

There is a global acceptance that communication is crucial to terrorism. Short Message Service (SMS), emails, instant and voice messages which are unstructured make up the top four most used means for terrorist communication which is largely based on natural language. In-order to detect evidence or threats conveyed in terrorist communication there is a need to analyse such forms of communication. To perform such analysis, we acknowledge that traditional text mining systems employ shallow parsing techniques and relies more on taxonomic relations or concept extractions which is not totally reliable in detecting semantic relationships and suspicious patterns in communication. Such task requires more complex analysis and processing of text. We propose an integrated framework that performs syntactic and semantic analysis of natural language. This integrated framework is developed using Natural Language Processing (NLP) techniques to process text, Ontology Based Information Extraction (OBIE) too understand, represent and express the problem domain, Computational linguistics technique(s); Phrase Structure Grammar (PSG, Context Free Grammar (CFG) and lastly, linguistic rules based on regular expressions to create a rule-based modelling of natural language from a computational perspective . Dataset was obtained from the Message Understanding Conference (MUC) and Global Terrorism Database (GTD) which consists of actual communication of terrorist activity. By analysing these datasets with the system, an average precision, recall, F-score of 90.5, 86.1, 88.43 and 95.5, 93.3, 92.63 for the MUC and GTD dataset respectively were obtained. The experimental result obtained clearly shows that our system not only identifies major conditions that satisfies the definition of a terrorist attack but also expresses the relationship, intent, recipient and location of an attack. This in turn informs the security analyst to take prompt decisions as regards such communication that includes malicious content.

Keywords- Ontology based Information Extraction, Phrase Structure Grammar, Context Free Grammar, Linguistic rules based on regular expressions.

I BACKGROUND

One of the major setbacks to development of a country is security. The concept of security has been central even in the primitive societies [21]. Security is a state of being safe with the absence of anxiety, fear, danger, poverty and oppression. Of the many threats to security, insurgency and terrorism have emerged as the most widely recognizable and visible threats to a nation's security especially after the 9/11 attacks and most recently the insurgency in Nigeria. Nigeria is the third of 62 most terrorized countries in the world just after Iraq and Pakistan [17].

Terrorist activities are increasing by the day. Global federal agencies and local Nigerian security agencies such as the State Security Services (SSS) are actively collecting domestic and foreign intelligence information to prevent future insurgency. Recently, the Global Terrorism Database (GTD) [20] released a concise report of terrorism with data and trends of attacks per country and individual attack type. Terrorist activity regardless of how clandestine they are, require communication of some sort. Communication via e-mail, short message service (SMS) have become an

Bolanle F. Oladejo and Noble I. Onyemenam (2019). Development of an Extended Natural Language Processing (NLP)-based Framework For Knowledge Discovery In Terrorism-based Communication, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 3, No. 1, pp. 60 - 71

essential way of exchanging information globally. Everybody uses e-mail, SMS and online chat applications as a form of interaction. These forms of communication are still one of the top three modes of terrorist communication with codes or the use of metaphors [11]. This communication media also consists largely of natural language in form of text or verbally.

The goal of Artificial Intelligence (AI) is to make machines do things that would ordinarily require intelligence if done by humans [3]. For a computer to be termed "intelligent" It would need to possess capabilities; (i) Knowledge representation to store what it knows or hears; (ii) Automated reasoning using stored information to answer questions and to draw new conclusions; (iii) Machine learning to adapt to new circumstances and to detect and extend patterns. (iv) Natural Language Processing to enable it to communicate successfully in English. Natural language processing (NLP) consists of using various techniques as well as Text Mining to extract underlying meaning in text [25]. Text mining can be referred to as the data analysis of textual resources for discovery of new, previously unknown knowledge [15].

The ability for computers to process human language is as old as the idea of computers themselves [18]. Natural Language Processing (NLP) is an interdisciplinary field of Artificial Intelligence and Linguistics, devoted to make computers understand the statements or words written in human languages [8]. NLP is a component of text mining that consists of techniques such as Information Extraction that can perform special kind of linguistic analysis that essentially helps a machine "read" text [24]. Information Extraction (IE) is concerned with locating specific pieces of data in natural-language documents, thereby extracting structured information from unstructured text [34]. The information is then stored in database like patterns and it can be available for further reuse. It converts a quality of textual document into more structured database [28]. NLP techniques aid Information Extraction [5, 6, 10, 13] such as Named entity recognition, tokenization, morphological and lexical processing, part-of-speech tagging all help in performing syntactic analysis of natural language. For example, part of speech tagging may aid in identification of entity types (verb: detonated, assassinated) or even targets or subjects (Noun: Civilian, people) or likely Noun phrases such as "terrorist group" or "armed bandits".

Due to the ambiguity of natural language, sentences can be long and complex resulting in a large number of patterns [35]. This limitation of NLP can be solved with Computational linguistics. Computational linguistics, introduced by Noam Chomsky in 1957 [4] is the application of linguistic theories and computational techniques to problems of natural language processing. Core techniques such as Phrase Structure Grammar (PSG) and Context Free Grammar (CFG) can be used to represent and reduce number of needed suspicious patterns in a sentence structure [35]. This work utilizes Phrase Structure Grammar (PSG) and CFG during syntactic analysis to reduce the number of patterns needed in IE rules. Reducing the number of rules is essential for making extraction rules more general and increases its extraction power.

To better understand the terrorism domain, literatures concerning the topic such as the Global Terrorism Database (GTD) [20] and Message Understanding Conferences (MUC) were studied. MUC's were introduced and financed by the Defence Advanced Research Project Agency (DARPA). Its aim was to encourage development of high performance Information Extraction systems through competition of research teams [14]. These repositories were essential in designing an ontology for terrorism for knowledge representation of the domain. This Section gives a contextual background of AI, Text Mining, Natural Language Processing (NLP), Information Extraction, computational linguistic techniques and ontology: Section II discusses the problem statement. In Section III, we reviewed recent related work advances in identifying communication with suspicious patterns or malicious content. Section IV and V describes the methodology and illustrates the operational phases and architecture of the proposed integrated framework to validate if such communication is malicious or not and steps for informing the security analyst. The experimental results when tested with dataset collected from the Message Understanding Conference (MUC 4) and the Global Terrorism Database (GTD) are shown in Section V.

Finally, Section VI concludes the paper with an outlook towards future research directions for improving the features of Terrorism Ontology Based Information Extraction.

II PROBLEM STATEMENT

Increase in terrorist activities globally shows security agencies are lacking in gathering useful information and data analysis [12]. Based on various threats and attacks that have been discussed in newspapers, television and unclassified news websites, terrorists often use e-mail, SMS, online chat which are unstructured to issue threats and to communicate the plan.

It is therefore imperative that effective and early detection of threat must entail active gathering and analysis of unstructured data. However, analysing and identifying evidence relevant to terrorism and suspicious activity from a large volume of unstructured data such as SMS, instant messages and email is challenging and time-consuming. Furthermore, traditional text mining which should be used for analysis employ shallow parsing techniques and focus on concept extraction and taxonomic relation extraction which is not adequate enough in detecting semantic relationships among word and phrases. This is further exacerbated by the absence of contextual expert knowledge of the problem domain.

III RELATED LITERATURE

The Text Mining technique using TF-IDF (Term Frequency-Inverse Document Frequency), Cosine similarity, Dice coefficient was implemented for detecting terror related comments on social networking sites [19]. This approach is however limited as the system counts frequency of words and does not consider sentence structure neither does it perform semantic analysis of natural language. Similarly, a knowledge engineering approach that makes use of lexicons, keyword in context (KWIC) index as well as rules for information extraction by analysing patterns of words in terrorism documents was proposed [30]. This system uses specific well known malicious words contained in the index to determine the structure of sentences, terms used together and number of times they occur. Such information determines the rules to be used in the information extraction process. This approach still does not consider semantic analysis and relationship between words or entities.

Data mining association classification rules was also implemented for detecting malicious content in instant messages [29]. The system makes use of an ontology that merely consists of a limited terrorist vocabulary which the system checks against the terms found in the instant message. Once again, semantic relationship of terms was not put into consideration in the ontology building. Association rule mining technique and basic ontology concepts was also adopted in the detection of terror communication in instant messages [31]. The framework detects suspicious messages from instant messaging systems in early stage and helps to identify and predict the type of cyber threat activity and traces the offender details. However, the system focuses on suspicious words (short and code forms) in its ontology without any proper relationship among key concepts. Another attempt at detecting the resonance of terrorist movement frames on web forums was carried out [10]. The system used a Terror Beliefs Ontology (TBO), Frame Discovery System (FDS) to capture and model various frames of terrorist behaviour and a Frame Resonance Detection System (FRDS). The system however does not perform semantic analysis so therefore does not address the problem of ambiguities in natural language.

IV METHODOLOGY

In this section, an extended Natural Language Processing (NLP) framework is developed to address the problem of detecting suspicious patterns in natural language-based communication with the intention of knowledge discovery in terrorism-based communication. To test the system, Dataset corpus from the MUC and GTD were chosen because they contain real terror attacks that describe all of the entities involved in each event as described in our developed ontology for terrorism.

ONTOLOGY DESIGN: The system is called Terrorism Ontology Based Information Extraction (TOBIE). Part of the preliminary processes is to first design an ontology for terrorism for knowledge representation for that domain. This work's ontology framework is an

extension of the Adversary – Intent – Target Model statement; a terrorist attack occurs when an adversary organisation, with intent and capability, uses a weapon against a target [32]. This model clearly requires definitions for words such as "attack," "adversary," "intent," "weapon," "target," etc. The following tasks are carried out in designing an ontology for terrorism;

1. Knowledge gathering and representation: This involved performing of structured and unstructured interviews and the study of existing literatures on terrorism.
2. Design of an ontology for terrorism: In the analysis and design of the ontology for terrorism, this work defines facts as; Terrorism occurs when an adversary using a Person/Agent, with Intent, uses a Weapon against a Target based on the **Adversary Intent Target** model [32]. An adaptation to this model for this work is the inclusion of the Weapon, Agent and location class because a target is found in a location. Likewise, any terrorist attack must have a location. The ontology for terrorism consists of the following classes;

    - Adversary Organisation: The organisation that masterminds terrorist attacks. This work makes use of Boko Haram, Al- Qaeda and IPOB (indigenous people of Biafra) as adversary organisations. This entity has object relationships of "**hasMember**" with Person/Agent and **"hasIntent"** with entity "Intent".
    - Person/Agent: This entity is a subclass of the supertype Adversary Organisation. It has a "**memberOf**" object relationship with its super type. Adversary Organisations do not carry out terrorist attacks directly. They only mastermind each attack. There is the need to differentiate Person and Agent. Most times, a terrorist's name may not be mentioned especially if it is not known. In this case, terms such as "assailants", "gun men" will be used and thus categorized as "Agent". This entity has object relationship of **"usesWeapon"** with entity "Weapon"
    - Weapon: Terrorism much often than not is carried out with a weapon which may be implicitly (code forms such as fireworks) or explicitly stated (such as bomb). Weapon has an object relationship of "**attacksWith**" with class Person/Agent.
    - Intent: Intents vary by organisation. This work considers only two intents – national separatist and religious extremism. For example, while Boko Haram or Al Qaeda's intent will be religious extremism, IPOB's intent is national separatist.
    - Target: This is the recipient of any terrorist action. It may be a public place, a person, group of people, private or public property. The object relationship between "Target" and "Location" entities is **"hasLocation"**
    - Location: This is the actual location a terrorist action occurs.

Table 1: Object Properties in Ontology for terrorism model.

Relation/Property	Domain	Range	Usage
hasMember	Thing	Adversary Organisation	Relationship between an attacker and the adversary organisation
hasTarget	Thing	Target	Relationship between the Organisation and target
attacksWith	Thing	Weapon	Relationship between weapon used in the attack and both adversary organization and member
hasLocation	Thing	Place	Of target or attack occurring or existing in some place.
hasIntent	Adversary Organisation	Intent	Relationship between an adversary organization and its intents.
memberOf	Person	Organisation	The relationship between a member and its organisation.
targetOf	Target	Thing	The inverse of hasTarget
Part of	Terrorism	Entities	Relationship between the classes Adversary Organisation, Target, Weapon and a Terrorist Attack

A. System Architecture for TOBIE

To detect suspicious patterns from communications based on natural language, this work adopts an approach for developing an extended NLP based integrated framework using text mining and information extraction techniques. In this Section we explore the operational phases of proposed integrated

framework called Terrorism Ontology based Information Extraction (TOBIE) as shown in Fig. 1. The approach utilizes syntactic Natural Language Processing (NLP) modelling and extended semantic techniques to facilitate automated analysis of textual and verbal terrorism related document processing for extracting semantic information elements from them in the form of annotations. This integrated framework examines the syntactic and semantic features of the text in defining these suspicious patterns and majorly employs the use of various pipeline processes for identification of ontology concepts and instances of suspicious patterns from messages and then extraction based on linguistic rules.

Phase I – Corpus Input: The system deals with the processing of natural language (speech, document, plain text, email, short message service). It assumes the input into the system as either verbal or plain text documents (pdf, html, txt, docx). In a case where the unstructured communication format is verbal, a transcription will be done to formally convert the speech into plain text representation. The main setback to speech to text transcription in the terrorism domain are:

- Almost all software has to be "trained" to a specific speaker to maximize accuracy. Also, the degree of accuracy depends on how slowly and distinctly the person speaks.
- Unavailability of real pre-recorded recordings of terrorism communication. In this case developmental speech recordings are created and utilised.

Phase II - Pre-processing: This phase prepares the raw (i.e. unprocessed) text for further syntactic analysis and processing. Pre-processing consists of tokenization, POS tagging, de-hyphenation, sentence splitting and morphological analysis [35].

Phase III- Feature Generation: This work utilizes; domain-specific ontology-based semantic features [26] in addition to syntactic features. The feature generation methodology consists of named entity recognition, gazetteer compiling and pattern generation.

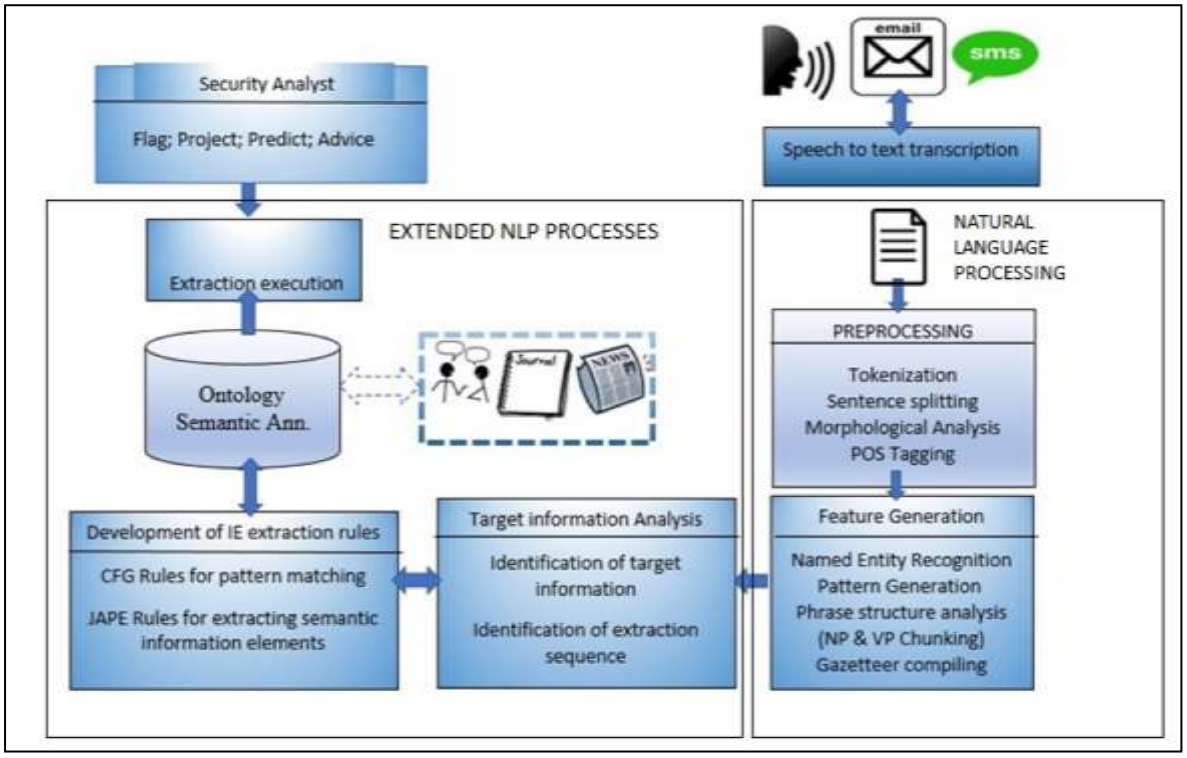

Fig. 1. Framework for Terrorism Ontology based Information Extraction

Phrase Structure Analysis (Noun and verb chunker)

- Named Entity Recognition: This stage extracts relevant Named Entities (NE) from the analysed document. NER can be described as the identification of words in text that correspond to a predefined taxonomy such as person (PER), organization (ORG), location (LOC), date and time [34]. NEs represent real world entities. In other words, Named Entities can be considered as instances of ontological concepts (e.g. grenade is an instance of Weapon or Osama Bin Laden a Person).
- Gazetteer Compiling: A gazetteer contains a set of lists containing names of specific entities (e.g. cities, organizations) [7]. Generally, a gazetteer list could group any set of terms based on any specific commonality possessed by these terms. The feature generation phase utilizes the information that a word or phrase belongs to a certain list in the gazetteer as a feature for IE tasks.
- Pattern Generation: This section elaborates the way context features are obtained. In order for TOBIE to extract information from communication documents, the system must know the syntactic structure of the document. First, the information obtained from the pre-processing phase and then the use of a parser. Parse trees are used to represent the syntactic structure of a given sentence. The Internal nodes represent the syntactic constituents of the sentence while its leaf nodes are the word tokens of the sentence [16]. The possible syntactic constituents are S(clause), VP (verb phrase), NP (noun phrase), PP (prepositional phrase), etc Fig. 2 shows a parse tree for a simple sentence; Boko Haram claimed responsibility

Syntactic patterns analysis is applied to the corpus so that suspicious patterns are generated for every lexical instantiation of the patterns that appears in the corpus. For example, the pattern "PassVP" would generate extraction patterns for all verbs that appear in the corpus in a passive construction or tense. In the terrorism domain, some of these extraction patterns might be: "PassVP (assassinated)" and "PassVP (terrorised)." These would match sentences such as: "the ambassador was assassinated", and "the embassy and hotel were bombed". PassVP refers to passive voice verb phrases (VPs), ActVP refers to active voice VPs, InfVP refers to infinitive VPs, and AuxVP refers to VPs where the main verb is a form of "to be" or "to have".

Phase IV – Target Information Analysis: This phase is for manually analysing the text to identify the types of semantic information elements to be extracted and their inter-relationships, and the sequence of their extraction. This consists of two phases:

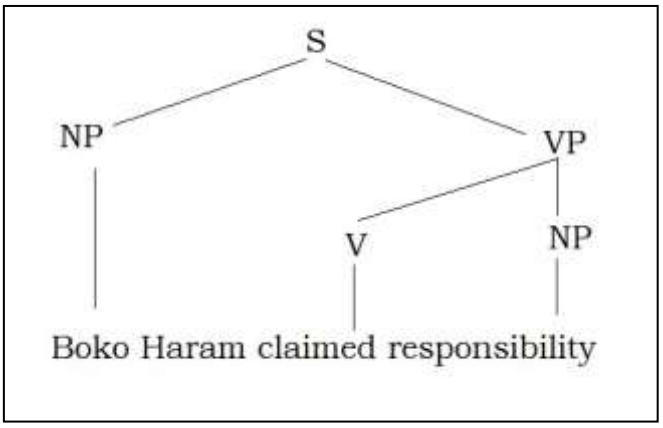

Fig 2. A Simple Sentence Structure

- Identification of Target Information: text is manually analysed to identify the types of semantic element requirements that are expressed in the text.
- Identification of Extraction Sequence: The purpose of this step is to identify the sequence of extracting the semantic information elements. Based on experimental studies, it has been found that extracting all semantic information elements from a sentence by a single IE rule (i.e. extracting all instances at the same time) is not efficient.

Phase V - Phrase Structural Grammar (PSG): Because of the ambiguity of Natural language [35], TOBIE needs to determine the general linguistic units. To do so, we need to do sentence analysis automatically and represent each word as constituent parts. Phrase structural analysis uses Phrase structure tags (determined by POS tagging), NP and VP chunkers as its basis, its job is to assign labels to each phrase of a sentence. In order to rewrite these labels or suspicious patterns as rules, we made use of Context Free Grammar (CFG). CFG describes possible strings in any given formal language. CFG grammar consists of a set of rewrite rules which associates a single nonterminal symbol with a string of terminal and nonterminal symbols [22].This rule was put into additional consideration when writing the Java annotated pattern engine (JAPE) rules. In order to correctly identify a weapon used in a malicious context, a verb (verb phrase) basically will have to precede the weapon in the text since verbs defines the particular action in which the weapon is being used. For example, the phrases, "Dear Uthsman", "you shall detonate the weapon" "a hybrid nuclear bomb in the church premises" are assigned NP, VP, and NP tags, respectively as seen in Fig. 3. PSG is employed to generate phrasal tags. Application-specific PSG rules are derived based on randomly selected sample of text called "development text", (which is also used for text analysis and further development of IE rules). Applying these PSG rules, phrasal tags are assigned when a certain combination of POS tags and/or phrasal tags are encountered. For example, the rule "Location" "IN DT NN" states that the phrasal tag "Location" should be assigned when the sequence of POS tags "IN DT NN" is encountered. Our use of phrasal tags together with PSG reduces the possible number of enumerations in patterns.

Phase VI - Ontological Knowledge Representation: The ontology conceptualises terrorism, this phase extracts named entities and defines the representation format for extracted information. this work utilizes the ACC-tuple to represent the extracted information. This is because it is easy for computer manipulation and thus evaluation (e.g. <Subject, Attribute, Value>). A six-tuple format is used for intermediate processing. In this representation, each element is called a "semantic information element". A semantic information element is an ontology concept or ontology relation. The six-tuple format for intermediate information representation: <AdversaryOrganisation, Weapon, Target, Location, Intent>.

- **Ontology-based semantic analysis**: The concept and relation of the ontology assists in extracting the semantic features of the text. A partial (and schematic) view of the ontology, including its concepts (e.g. adversary organisation) and sub-concepts (e.g. intent).
- **Semantic Annotation:** The purpose of semantic annotation is to match features with the appropriate ontology classes.Semantic annotation is a process of adding semantic information to linguistic forms. From two set of objects, document and formal representation; a function can be defined. The function that maps documents to formal representation, called annotation as seen in fig. 4. Semantic matching is further sub-divided in two; Discovering subsume concepts and Ontology matching.

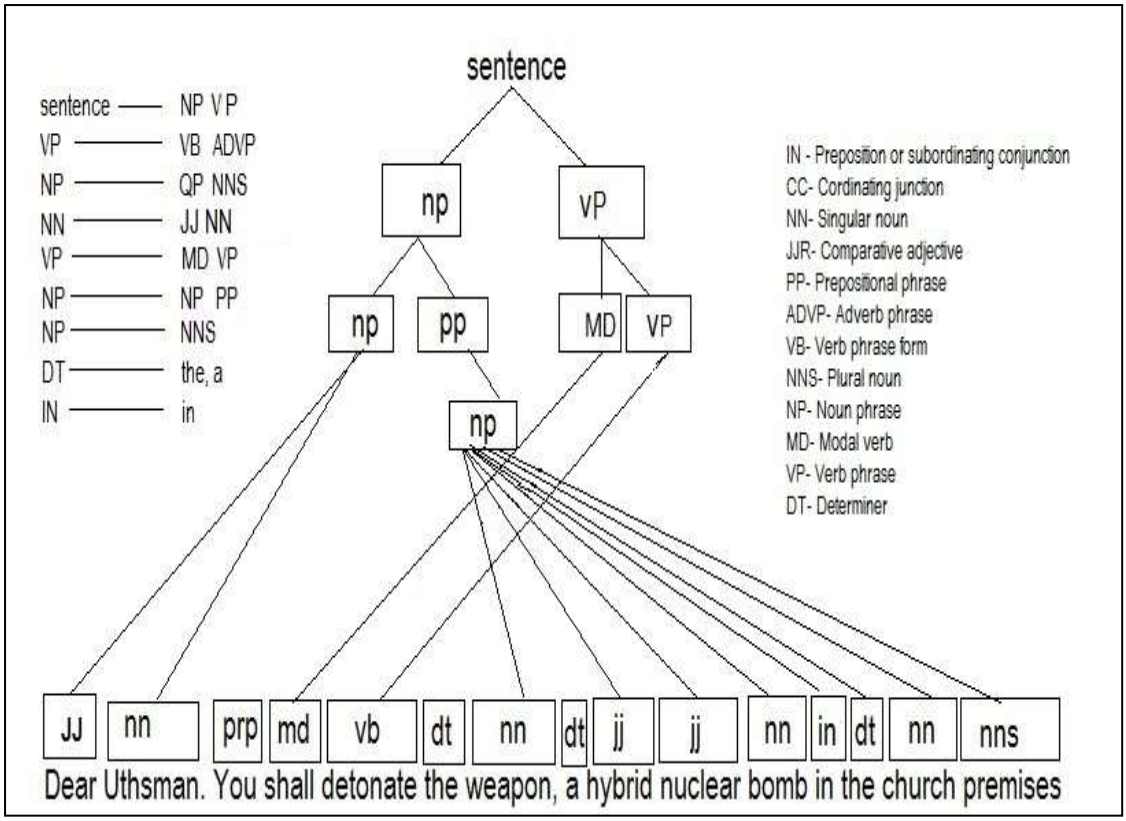

FIG. 3: CFG Rules for Developmental Text

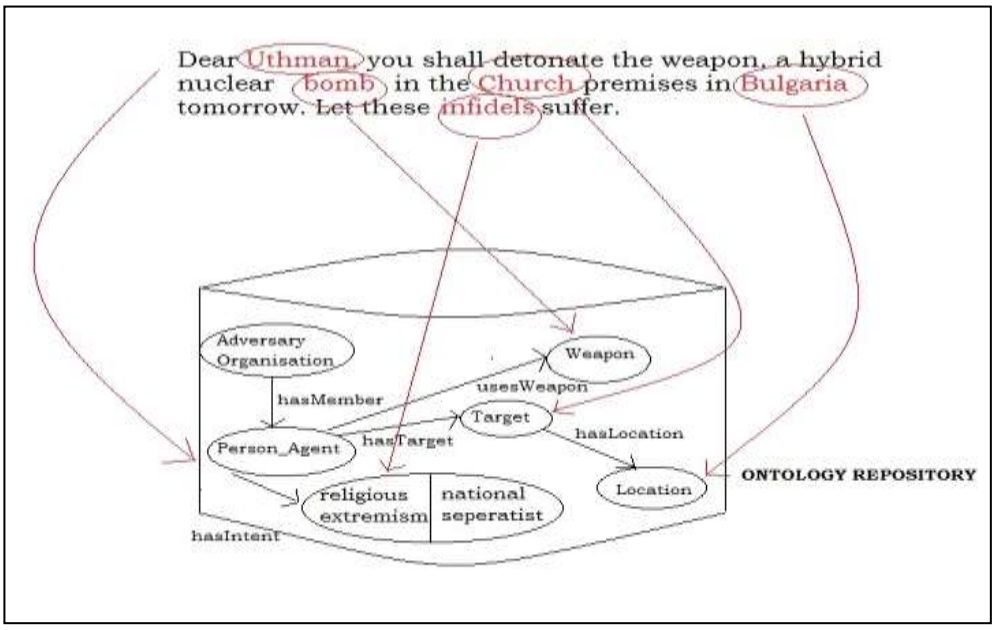

FIG. 4: Semantic Annotation

Phase VII – Development of Information Extraction Rules: This phase develops a set of grammar rules which automatically executes the information extraction process. This work utilises rules for extracting single semantic information elements (IE rules). The extraction IE rules are based on pattern matching methods. The left-hand side of the rule defines the pattern to be matched while the right-hand side defines

which part of the matched pattern will be extracted.

Syntactic (POS tags, NP VP chunker, gazetteer terms etc) and semantic (ontology concepts and rules for extracting single semantic information elements (IE rules).

The extraction IE rules are based on pattern matching methods. The left- hand side of the rule defines the pattern to be matched while the right-hand side defines which part of the matched pattern will be extracted. syntactic (POS tags, NP VP chunker, gazetteer terms etc) and semantic (ontology concepts and relations) text features found in the patterns of the IE rules are utilized. If a concept in the ontology is used in the IE rule, all its sub-concepts will be included in the matching as well. For example, in the following IE rule, "weapon" is a concept in the ontology: Once "weapon" is matched, it will extract the matched text as an instance for "weapon". Applying this IE rule "bomb" will be extracted as an instance of "weapon" because it matches a sub-concept of "weapon" in the ontology.

Phase VIII - Extraction Execution: This phase aims at extracting the target information element instances from the terrorism related documents text using the rules developed in Phase IX.

Phase IX– Security Analysis: the system provides facts based on the rules developed in phase VIII. The extracted suspicious patterns as represented with annotations of the natural language corpus are presented to a security analyst to make efficient use of these facts. It will be the duty of the analyst to thereby flag, make projections or perform further analysis on these patterns.

V. EXPERIMENTAL RESULTS

A. Dataset

Due to the unavailability of real suspicious content online, this work used a corpus consisting of 200 news selected from the corpus TST3 and TST4 of the 4th Message Understanding Conference (MUC4). These news articles are on terrorist activities in Latin American countries from the late 1980's to early 2000s and contain a hundred articles with over 6000 words. The second category of corpus (GTD) [20] was chosen because it contains records of terrorist activities in Nigeria. The corpus was chosen because they contain templates that describe all of the entities involved in each event as described in our developed ontology for terrorism. An example snippet of attacks from the GTD and MUC-4 corpus is as follows:

Assailants *detonated an* ***explosive device*** *at a Joint Task Force (JTF)* ***checkpoint*** *in* ***Maiduguri city, Borno state, Nigeria****. At least one* ***soldier*** *and 13 assailants were killed in the ensuing firefight; at least* ***two people*** *were also injured. Officials attribute the attack to* ***Boko Haram.***

The ***terrorists*** *used* ***explosives*** *against the* ***town hall****. El Comercio reported that alleged* ***Shining Path members*** *also attacked* ***public facilities*** *in* ***huarpacha****,* ***Ambo****,* ***tomayquichua****, and* ***kichki****. Municipal official* ***Sergio Horna*** *was seriously wounded in an* ***explosion*** *in Ambo.*

B. Evaluation Method for Dataset

For evaluation of semantically annotated terms from dataset, IE performance metrics of precision, recall and F-score was used. They are calculated as:

$$Precision = \frac{|\text{correct answers}|}{|\text{total answers}|}$$

Precision measures reliability and accuracy by determining what percentage of the information extracted is correct [33].

$$Recall = \frac{|\text{correct answers}|}{|\text{correct answers in the gold standard}|}$$

Recall measures the percentage of available correct information extracted, thus measuring the ability of the system to extract relevant information [22].

$$\text{F - measure} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

F-score is a combined weighted measure of recall and precision. It combines recall and

precision by using the harmonic mean of precision and recall [2].

C. Results Obtained

The outputs obtained from the evaluation of TOBIE using MUC4 and GTD dataset are shown in Tables II and III respectively. TOBIE achieved an average precision, recall and F-score of 90.5, 86.1 and 88.43 for the MUC 4 dataset and 95.5, 93.3, 92.63 Precision, recall and F-score for the GTD dataset. Recall that for this work, the ontology for terrorism was designed considering only Boko Haram, Al-Qaeda and IPOB as adversary organisations along with their corresponding intent. However, these adversary organisations are not rampant in Latin America as the MUC4 dataset suggests. This explains why TOBIE had slight difficulty in recognising such adversary organisation such as Shining Path Members is peculiar to Latin America.

VI. RECOMMENDATION / FUTURE DIRECTION

Despite the high performance TOBIE achieved, three limitations of this work are acknowledged. This should be addressed as part of future/ongoing research:

1. We cannot rule out the possibility of messages being encrypted. In such cases, an appropriate decryption technique should be used [9].
2. Ontology adaptation for Multi-lingual languages can be included in the ontology [1].
3. Third, we only tested our methodology/algorithms on terrorism-based communication. In future work, we will extend our methodology/algorithms to extract information from other types of documents (e.g. job vacancies, environmental regulatory documents), as well as contractual documents (e.g. contract specifications).

Table II: Result Of Extraction Method for MUC-4 dataset

Named Entity	Nos of facts	Correctly identified	Total identified	Precision	Recall	F-score
Person Agent	69	63	67	94	91	92.47
Adv. Organisation	58	50	54	92	86	89.90
Intent	56	40	51	78	71	74.33
Weapon	73	65	69	94	89	91.43
Target	75	67	73	91	89	89.98
Location	60	55	58	94	91	92.47

Table III: Result of Extraction Method for GTD Dataset

Named Entity	Nos of facts	Correctly identified	Total identified	Precision	Recall	F-score
Person Agent	62	58	61	95	93	93.98
Adv. Organisation	60	56	58	96	93	94.47
Intent	60	56	59	94	93	93.49
Weapon	79	75	78	96	94	91.43
Target	87	81	85	95	93	89.98
Location	89	84	86	97	94	92.47

VI. IMPLICATION TO RESEARCH AND SOCIETY

This work advances research from two perspectives, intellectually and application wise. Intellectually, it advances knowledge in three major ways. First, it provides an efficient integrated framework that benefits from expert domain knowledge designed in an ontology and NLP coded in the form of IE grammar rules. It also shows that the efficiency of algorithmic development for rule-based methods can be enhanced through the use of PSG-based phrasal tags. Third, it proves that efficient NLP can be successfully achieved if supported with domain knowledge (represented in the form of a domain ontology) and expert NLP (represented in the form of IE rules) are captured and integrated in one platform.

From the application perspective, the application of this work can be extended to support automated information extraction and analysis for other applications and purposes such as analysis of job adverts, contract document analysis for inconsistencies etc.

VII. CONCLUSION

Previously, there have been numerous attempts at information extraction from documents and news articles in natural language. However, such attempts do not consider sentence structure neither do they address the problem of ambiguity in natural language. Furthermore, Information Extraction should be carried out while guided by an ontology. This gives the system background knowledge or understanding of how each class or object found in the communication media is related as the terrorism domain suggests. This work employs Information Extraction techniques on Nigerian reports on Terrorism. Unlike many systems that use Machine learning techniques, the system, TOBIE, adopts a rule-based knowledge engineering approach that relies on syntactic and semantic analysis. This approach allows TOBIE to use the syntactic and semantic structure of language in analysing communication, rather than using machine learning for training, testing or other statistical analysis.

REFERENCES

[1] Ali M. M., and Lakshmi Rajamani, "Framework for Surveillance of Emails to Detect Multilingual Spam and Suspicious Messages," IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions , IIT Kanpur, India, pp. 42-56, 2013.

[2] Appelt, D. E. "Introduction to information extraction," AI Communications (12:3), 1999, 161-172.

[3] Boden, M. F. (1977). Artificial intelligence and natural man. Publisher, Basic Books, ***1977***. ISBN, 0465004539, 9780465004539

[4] Chomsky, Noam, 1965, Aspects of the Theory of Syntax, Cambridge, Massachusetts:MIT Press.

[5] Cowie, J., and Wilks, Y. "Information Extraction," inHandbook of Natural Language Processing. , R. Dale, H.Moisl and H. Somers (eds.), Marcel Dekke: New York, 2000,241-260.

[6] Craven, M., DiPasquo, D., Freitag, D., McCallum, A.,Mitchell, T., Nigam, K., and Slattery, S. "Learning toconstruct knowledge bases from the World Wide Web,"Artificial Intelligence (118:1-2), 2000, 69-113.

[7] Cunningham H., Maynard D., Bontcheva K. and Tablan V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).

[8] Diksha Khurana, Aditya Koli, Kiran Khatter, Sukhdev Singh (2017) Natural Language Processing: State of The Art, Current Trends and Challenges. Department of Computer Science and Engineering. Manav Rachna International University, Faridabad-121004, India.

[9] Thambiraja, E. G. Ramesh, and Uma Rani, "A Survey on Various Most Common Encryption Techniques," published by IJARCSSE Journal volume 2 issue 7, pp. 226-233, 2012.

[10] Etudo U. (2017) Automatically Detecting the Resonance of Terrorist Movement Frames on the Web.scholarscompass.vcu.edu/etd/4926/

[11]Gaizauskas, R., and Wilks, Y. "Information extraction:Beyond document retrieval," Journal of Documentation(54:1), 1998, 70-105.

[12] Gardner, F. (2013). How do terrorists communicate? http://www.bbc.com/news/world-24784756.

[13] Gowri S., Anandha G. S. and Divya G. (2014), Suspicious data mining from chat and email data, International Journal of Advances in Science Engineering and Technology, ISSN: 2321-9009 Volume- 2, Issue-2.

[14] Grishman, R. "Information extraction: Techniques andchallenges," in Information Extraction A MultidisciplinaryApproach to an Emerging Information Technology, Springer,1997, 10-27.

[15] Grishman, R., and Sundheim, B. "Message Understanding Conference-6: A Brief History." In COLING-96, Copenhagen, Denmark, 1996, 466-471.

[16] Hearst, M. (2003) "what is text mining" SIMS,UC Berkeley.

[17] Hsinchun Chen, Richard Miranda, Daniel D. Zeng (2003) Intelligence and Security Informatics: First NSF/NIJ Symposium, ISI 2003

[18] Institute for economics & peace: Global terrorism index 2015.
[19] Jurafsky, D and Martin, J. H. (2008). Speech and Language Processing an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.
[20] Kulkarni S., Sayali, B. P. and Ganesh S.(2016) Terror Detection Using Text Mining. Imperial Journal of Interdisciplinary Research (IJIR)Vol-2, Issue-5, 2016ISSN: 2454-1362.
[21] National consortium for the study of terrorism and responses to terrorism (START) (2017). Global Terrorism Database [Data file] Retrieved from https://www.start.umd.edu/gtd.
[22] Nwanegbo. C. J. andOdigbo, J.(2013), Security and National Development in Nigeria: The Threat of Boko Haram International Journal of Humanities and Social Science Vol. 3 No. 4.
[23] Owen Rambow and Aravind Joshi (1994) A Formal Look at Dependency Grammars and Phrase-Structure Grammars, with Special Consideration of Word-Order Phenomena. n Leo Wanner, ed., Current Issues in Meaning-Text Theory. Pinter, London, 1994
[24] Raymond J. Mooney and Razvan Bunescu (2005) Mining Knowledge from Text Using Information Extraction. Department of Computer SciencesUniversity of Texas at Austin SIGKDD Explorations. Volume 7, Issue 1 - Page 3.
[25] Robin, T. S. (2012). Natural Language Processing: Parts-of-speech tagging, [Online]. Available:http://language.worldofcomputing.net/pos tagging/parts-of-speech-tagging.html
[26] Russell, S. J., Norvig, P., & Davis, E. (2010). Artificial intelligence: a modern approach. 3rd ed. Upper Saddle River, NJ: Prentice Hall.
[27] Saidah Saad, MohanaadAlnaseeri (2016) Domain-specific ontology-based approach for Arabic question answering. Journal of Theoretical and Applied Information Technology 1083(1):43 · February 2016
[28] Sagayam, R., Srinivasan, S., Roshni S. (2012). A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques, International Journal Of Computational Engineering Research (ijceronline.com) Vol. 2 Issue. 5, ISSN 2250-3005.
[29] Shendurkar A. and Chopde N. (2015) An Ontology based Enhanced Framework for Instant Messages Filtering forDetection of Cyber Crimes. International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169Volume: 3 Issue: 5.
[30] Sumali J. Conlon, Alan S. Abrahams and Lakisha L. Simmons (2015) Terrorism Information Extraction from OnlineReports. Journal of Computer Information Systems, 55:3,20-28.
[31] Thivya.G. and ShilpaG. (2015) Survey on Vigilance of Instant Messages inSocial Networks Using Text MiningTechniques and Ontology. International Journal of Innovative Research in Computerand Communication Engineering. Vol. 3, Issue 2, February 2015.
[32] Turner, M. D., Weinberg, D. M. and Turner J. A. (2011). A Simple Ontology for the Analysis of Terrorist Attacks, University of New Mexico Electrical and Computer Engineering Department Technical Report EECE-TR-11-0007.
[33] Van Rijsbergen, C. J. Information retrieval (2nd ed.).London: Butterworths, 1979.
[34] Zamin N., Oxley A. and Abubakar, Z. (2013). projecting named entity tags from a resource richlanguage to a resource poor language. Journal of ICT, pp: 121–146
[35] Zhang, J., and El-Gohary, N.M. (2012b). "Extraction of construction regulatory requirements from textual documents using natural language processing techniques." Proc., 2012 ASCE Intl. Conf. on Comput. Civ. Eng., ASCE, Reston, VA, 453-460.