TERRORISM INFORMATION EXTRACTION FROM ONLINE REPORTS

SUMALI J. CONLON
University of Mississippi
University, MS 38677

ALAN S. ABRAHAMS
Virginia Tech
Blacksburg, VA 24061

LAKISHA L. SIMMONS
Belmont University
Nashville, TN 37212

ABSTRACT

Many documents containing information about intelligence and security issues are available both in printed and electronic formats. In this research, we built an experimental system to extract intelligence and security information from electronic documents. Our system, CAINES, is based on a knowledge engineering approach and relies on sublanguage analysis techniques. CAINES performs syntactic and semantic analysis and uses lexicons of various categories of terms. The system is able to extract certain types of information from reports on terrorist incidents posted by the National Counterterrorism Center (NCTC), such as what happened and the results of the incidents.

Keywords: Information extraction, Terrorist incident extraction, Lexicon-Based analysis, Sublanguage analysis

INTRODUCTION

The tremendous growth in the number of online documents has created opportunities for people to gain enormous amounts of information electronically, and can also be used by computers for various applications. One of the main difficulties in dealing with online documents, however, is that most of these documents are in textual form and require significant human effort to input the data into databases for analysis and reporting. This process requires a great deal of time, and some information can be missed if the documents are large. Thus, in this research, we built a system that semi-automatically analyzes the contents in documents containing intelligence and security information published by the National Counterterrorism Center (NCTC). The goal of this research is to extract rapidly, and with high fidelity, from terrorist incident report narratives, structured information that can be tabulated into a database. This paper describes and evaluates a method for rapid, accurate, and complete information extraction from such textual narratives. This information extraction task is important as it could facilitate the creation of highly useful multi-dimensional online analytical processing (OLAP) reports [6] – such as "incidents by time of day", "incidents by week of year", "incidents by place", "incidents by mode of attack", "terrorist organization by type of weapon used", and others – that could be of significant managerial value to commanders and planners in the defense services.

Our system is based on a knowledge engineering approach rather than the fully automatic statistical technique used in machine learning methods [46]. As described by Appelt [3] the knowledge engineering approach performs well when linguistic resources such as lexicons are available. We built our lexicons and knowledge bases by analyzing data from the collection of reports published by NCTC from year 2001 to 2012. Syntactic and semantic analyses are also used to help in the information extraction process. The process is referred to as "semi-automatic" as some initial effort is required to construct the lexicons (e.g. list of place names, list of person names, list of organization names), prior to information extraction. While the lexicon creation and information extraction steps can be mostly automated, some human cross-checking and correction effort is helpful to improve accuracy and completeness.

We analyze the contents of these electronic text documents automatically using our information extraction system CAINES (Content Analyzer and Information Extraction System). CAINES has been used to analyze financial information from online news articles published by several sources such as Reuters and NASDAQ [9]. It has also been applied to the analysis of online product and movie reviews [67]. We use the "knowledge engineering" approach, with assistance from artificial intelligence and linguistic analysis techniques to build systems to analyze texts in specific domains.

The rest of the paper is organized as follows. We begin with an introductory background to the Information Extraction (IE) field, together with its history and prior IE tools and applications, and we contrast these to the CAINES system. Next, we describe our methodology: our data set, tools and architecture, keyword in context (KWIC) index, and lexicons. We then describe our information extraction process, and provide pseudo-code procedures, as well as sample inputs and outputs of this process. Finally, we describe the results of a substantive evaluation of the accuracy and comprehensiveness of the system on a sample of almost one thousand pages from our source data set. We conclude with a summary and goals for future work.

BACKGROUND

Information extraction is defined as "the process of selectively structuring and combining data that are explicitly stated or implied in one or more natural language documents" [51]. Alternatively stated, Information extraction refers to "the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources" [63]. Information extraction is sometimes referred to as template filling, slot filling, or semantic frame filling. Generally, this refers to the conversion of free-form natural language text data into a more structured representation, where instances of entity types are identified, classified, and related to each other (e.g. via events or predicates).

A number of computational linguistics tasks may facilitate information extraction [10, 12, 13, 25, 29], such as tokenization, morphological and lexical processing, word sense tagging, part-of-speech tagging, event and relation extraction, and co-reference resolution. We briefly review each of these tasks. Tokenization involves splitting paragraphs into sentences, and sentences into words. An example of morphological and lexical processing is the stemming of words like "killing" and "killed" to the root form "kill". Word sense tagging is necessary, for instance, to recognize whether "arrested" is used in the sense of "capture", rather than in the sense of "slow down". Part-of-speech tagging involves identifying, inter alia, nouns and verbs, and finding the subjects and objects of the verbs. Part-of-speech tagging may facilitate

event and relation extraction since it allows the determination of the event (verb: "attack") and the subjects and objects - e.g. "Zarqawi's group" (noun phrase) attacked "the Jordanian embassy" (noun phrase). While events are often referred to with verbs ("The group *attacked*…"), they may sometimes be referred to with nouns, so event extraction can be more complex than extracting verbs and their subjects and objects. Consider "Al Zarqawi's group *conducted* an attack when it bombed the Jordanian" or "The group *carried out* their last attack in August", where "conducted" and "carried out" are the active verbs, but the event is referred to by the noun "attack" (and specifically, the attack is a bombing, in the first case). Finally, co-reference resolution is necessary to pinpoint the identity of the entity being referred to via a pronoun or indirect reference, for example, determining that in "Last month, *the group* conducted an attack", the reference "the group" refers to "Zarqawi's group" introduced elsewhere.

A number of common concept types are typically extracted with information extraction systems. Named entities may be classified, for instance, as persons (e.g. "Nicholas Berg"), organizations (e.g. "Al Qaeda"), or other types of objects – a "VBIED" is a type of bomb, an "M15" is a type of gun, "daggers" and "axes" are types of knife weapons. Information extraction may extract actions ("killed") or states ("was hurt"), and each of these may be classified by type: "murdering", "slaughtering", "beheading", "fatally wounding" are ways of "killing"; "bombing", "striking", "shooting", and "hitting", are ways of "attacking". Semantic roles, such as agent, experiencer, cause, patient, source, location, and manner may be extracted from sentences [27, 55]. In terrorism, these semantic roles may correspond, for instance, with the perpetrator, victim, and instrument or weapon. Information extraction may extract time references such as the date (day, month, and year), time of day, or day of week. Similarly, various types of locations may be extracted (country, province, city, suburb). Measures and units of measure may be extracted (e.g. "*5 ton* bomb", "*three attacks*").

As the variety of forms of expression is so extensive, information extraction is a complex and frequently inaccurate task. However, while perfectly accurate automated extraction is elusive, partial success may nevertheless be valuable and helpful. Unstructured information processing often involves a trade-off between achieving high volume (using partially-accurate automated information extraction) and achieving high accuracy (using highly-accurate but slower human reading and annotation).

Next we turn to a review of the history of information extraction, and prior information extraction tools and applications.

HISTORY OF INFORMATION EXTRACTION

The incorporation of syntactic and semantic information in Natural Language Processing (NLP) techniques brought about new ideas in IE in the 1960's and 1970's [73]. In 1970, Naomi Sager and her colleagues first successfully applied IE technology to extract hospital discharge information from patient records. Dictionary lookup and pattern matching were used to extract relevant medical information [61]. In the early 1970s, Gerald DeJong developed an IE system named FRUMP. FRUMP used keywords and sentence analysis on newswire articles to determine the relevant information [10]. FRUMP became the basis of an IE system named ATRANS, the earliest IE system to be used for commercial purposes. ATRANS was able to extract bank money transfer information from telexes by using sentence analysis. The assumption of ATRANS was that the sentence structures of money transferring telex messages were predictable [44].

Message Understanding Conferences (MUC), Text REtrieval Conferences (TREC), and the TIPSTER text program helped define and promote research in IE. MUC's were initiated and financed by the Defense Advanced Research Project Agency (DARPA). DARPA is the research and development arm of the U.S. Department of Defense and often supports research and technology. The purpose of the MUC was to encourage development of high performance IE systems through competition of research teams [31]. MUC organized 7 different conferences (MUC-1 in 1987 to MUC-7 in 1997). Prior to each conference, MUC invited different research teams to develop IE systems based on the same data set. In order to develop unbiased systems, MUC provided the test data set one month before the conference.

MUC-1 (1987) and MUC-2 (1989), focused on extracting information from short naval messages. Many of the first systems that analyzed natural language text-based information came from MUC-1 and MUC-2 [36]. MUC-3 (1991) and MUC-4 (1992) centered on systems that extracted data about terrorists in Latin America from newspaper and newswire articles. The conferences continued in 1993, 1995 and 1997 (MUC-5, MUC-6, and MUC-7) using news articles to extract information about joint ventures, space vehicles, and missile launches [3].

INFORMATION EXTRACTION TOOLS AND APPLICATIONS

A considerable number of information extraction tools have been developed for various applications. One of the earliest such tools for financial applications was JASPER (Journalist's Assistant for Preparing Earnings Report), developed by the Carnegie Group of Reuters Ltd., which was designed to extract corporate news stories related to earnings, dividends, or income [2]. JASPER required manual evaluation using a set of test documents collected from PR Newswire. JASPER employed knowledge representation, syntactic and semantic knowledge of sentences, and domain dependent regularities of patterns, with higher precision of extraction than previously developed systems.

FIRST (Flexible Information extRaction SysTem) was able to extract specific information such as "sales rose 5%" from The Wall Street Journal (WSJ) [9]. FIRST built a knowledge base by using a training set of documents from the WSJ and relied on a service-oriented framework with information retrieval (IR) and IE components. The IR component retrieved source documents and the IE component analyzed the documents and converted them into a data template.

CAINES (Content Analyzer and INformation Extraction System) was built using a knowledge engineering approach to analyze documents from online sources including web blogs, customer reviews, government reports, and online news articles. CAINES analyzes texts using syntactic and semantic techniques. In 2009, CAINES was used to analyze electronic word-of-mouth (eWOM) reviews from a sample of 18 action and adventure movies consisting of 20,677 individual reviews [67]. Results from the study showed that storyline is most important and consumers tend to leave more positive than negative reviews. Findings also revealed key sentiments of consumers' evaluations towards movies, something not found in many other studies.

Some other notable previous IE research prototypes include SCISOR [36], EDGAR-Analyzer [26], Edgar Extraction System [28], Edgar2xml [40], and others. Table 1 provides a sample of historic information extraction tools. For a review of these, and others such as LIEP, PALKA, HASTEN, DiscoTEX, SRV,

TABLE 1. Sample of Historic Information Extraction Tools

SCISOR [36]	Diderot [11]
FASTUS [34]	WHISK [69]
GATE [12]	Web - > KB [13]
(LP)2 [8]	RoadRunner [14]
KIM [54]	Autoslog / Sundance [60]
ANNIE [15]	ReVerb [20]
TextRunner [4]	SEMAFOR [16]

CRYSTAL+Webfoot, and RAPIER, see [5, 19, 38, 52].

Due to language complexity, developing a general-purpose information extraction tool has proved elusive, and authors typically focus on narrow domains with more restricted lexicons. Table 2 summarizes historic applications of information extraction in narrow domains.

Table 2. Sample of Historic Application Domains for Narrow-Purpose Information Extraction Tools

Application Area (Domain)	Authors
News tracking	[36]
Joint ventures	[10, 11]
Academic citations	[17]
Business intelligence gathering	[62]
Microelectronics	[12]
Project reports	[65]
Corporate news	[2, 9]
Corporate filings	[26, 28, 40, 66, 68]
Terrorism	[3, 29, 39, 56, 57, 58]
Customer care	
Data cleaning	See [32, 47, 59, 63]
Classified advertisements	
Comparison shopping	

In the area of terrorism, various authors have attempted to extract incident information, particularly as it relates to terrorists in Latin America, from newspaper and newswire articles [3, 29, 39, 56, 57, 58]. No prior studies, to our knowledge, have applied information extraction specifically to the US Department of State *Country Reports on Terrorism*, as is proposed here.

METHODOLOGY

Data

In August of 2004, the President of the United States implemented by executive order, the creation of a National Counterterrorism Center (NCTC) for all-source analysis of global terrorism. NCTC data is included in the Department of State's annual report, *Country Reports on Terrorism*. Yearly reports include updates on the role foreign countries have played on world terrorism, state sponsors of terrorism, terrorist safe havens, terrorist organizations, and terrorism deaths, injuries, and kidnappings of private U.S. Citizens.

The United States Department of State publishes annual reports on terrorism for countries meeting criteria set forth by legislation in Title 22 of the US Code, Section 2656f (the "Act"). Full Country Reports on Terrorism are due to congress by April 30 of each year and more importantly, are made available to the public at http://www.state.gov/j/ct/rls/crt/index.htm.

This United States Department of State *Country Reports on Terrorism* webpage (http://www.state.gov/j/ct/rls/crt/index.htm) contains links to PDF files from 2001 to 2012, and each report is at least 250 pages long, meaning that manual consultation and summarization would be extremely arduous. A portion of the 2007 report is shown in Table 3.

Table 3. A Portion of the Country Reports on Terrorism 2007

On July 2, Abdu Mohamad Sad Ahmad Reheqa drove a suicide vehicle-borne improvised explosive device (SVBIED) into a group of western tourists in Marib, killing him and several others. AQY claimed responsibility. Three days later, U.S.-trained Yemeni security forces killed the suspected leader of the SVBIED bombing, Ahmed Basyouni Dwedar, an Egyptian wanted in Egypt for his ties to the Muslim Brotherhood. On August 8 and 13, Yemeni security forces raided two houses, arresting 17 and killing four AQ-affiliated suspects while suffering one casualty.

Tools and System Architecture

The major tools we use in this research are the programming language Perl and the relational database management system, MySQL. Perl (Practical Extraction and Report Language) is a high-level, general-purpose, interpreted, dynamic programming language. It combines features from programming languages such as C and several UNIX tools that make Perl very flexible and adaptable to many other tools and programming languages. A collection of Perl modules are also available for download at CPAN (the Comprehensive Perl Archive Network - http://www.cpan.org/).

To develop the system, we looked at the documents in the domain covered by the research. We analyzed the terms in the documents collected and constructed a set of lexicons, as well as rules for information extraction, by analyzing patterns of words in the documents. The tools and techniques we use are based on ideas from the areas of natural language processing and information retrieval. The overall architecture of our system is shown in Figure 1.

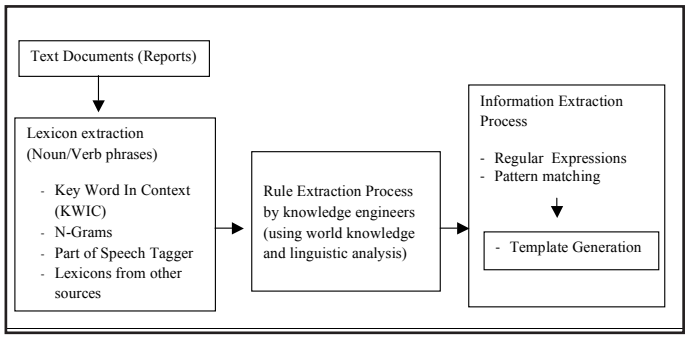


FIGURE 1. System architecture

The following sub-sections discuss each of these sub-systems:

KeyWord In Context (KWIC) Index:

A KWIC index is a file which is created by putting each word into a field in a row of a database [43]. After that, the first word is removed and each remaining word is shifted one field to the left and placed in the next row. The process continues until the last word in the sentence is put into the first position. For example, the sentence "The Islamic State of Iraq (ISI) claimed responsibility," can be entered in the database as shown in Table 4.

TABLE 4. KWIC Index Entries for the Statement "The Islamic State of Iraq (ISI) claimed responsibility"

The	Islamic	State	of	Iraq	(ISI)	claimed	responsibility
Islamic	State	of	Iraq	(ISI)	claimed	responsibility	
State	of	Iraq	(ISI)	claimed	responsibility		
of	Iraq	(ISI)	claimed	responsibility			
Iraq	(ISI)	claimed	responsibility				
(ISI)	claimed	responsibility					

We ran the KWIC index builder on eight reports from 2001 to 2012. Each report contained a file that consisted of about 100,000 records. We sorted these records on various columns to learn about the structure of sentences in these reports and to see which word combinations were frequently used. Some sample entries from the reports in 2009, containing the word "attacks" in column 4, are shown in Table 5.

TABLE 5. Sample Entries in the KWIC Index File Containing the Term "attack" in Column 4 from the 2009 Report.

Department	considers	terrorist	attacks	that	a	group	has	carried
would	launch	terrorist	attacks	throughout	the	world.	Following	the
probably	uses	such	attacks	to	discourage	foreign	investment.	The
No	major	terrorist	attacks	took	place	in	Western	Europe
Although	no	terrorist	attacks	took	place	on	French	soil
Most	of	the	attacks	took	place	late	at	night
assets	and	routinely	attacks	U.S.,	Coalition,	and	Government	of
AQI	claimed	its	attacks	under	the	MSC	until	mid-October,
terrorist	attacks	and	attacks	undertaken	by	â€œnational	liberation	movementsâ€&
(IEDs)	and	coordinated	attacks	usin	multiple	suicide	bombers,	resulting
cocktails.	Since	these	attacks	usually	occurred	outside	normal	business
large	scale	terrorist	attacks	were	carried	out	on	Uzbekistanâ€™s
vigilnce,	several	planned	attacks	were	disrupted	prior	to	execution.

The items in the KWIC index files can help us learn many things about the structure of sentences in the reports, such as which terms are used together and how many times. We use information in these files to determine rules to be used in the information extraction process. Details will be discussed in later sections.

The Lexicons

In order for a system to do language analysis well, it needs information about the words and phrases that appear in the reports as well as general knowledge. A lexicon is a dictionary of all the words in the language, which may contain many types of information about each word, for example, what part of speech it is (its lexical category), and what its distributional properties are [64]. The lexicon is used to store this word and phrase information while the knowledge base is used to store common sense knowledge. Some information about knowledge and semantic networks can be found in [71].

Since this research is related to intelligence and security studies, one of our lexicons is a lexicon that contains country names. This lexicon allows our system to recognize the country names in the reports. In the collection of reports from 2005 to 2008, the number of times different country names appear are shown in Table 6.

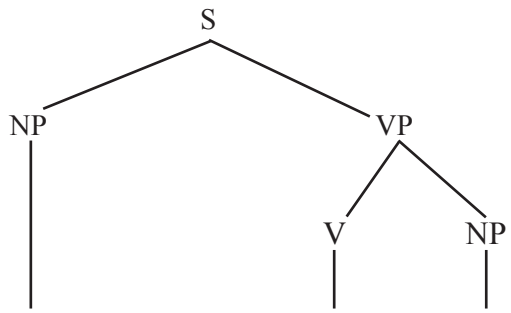
TABLE 6. Some Sample Number of Times Country Names Appear in Reports from 2005-2008.

Iraq	445	Somalia	8
Afghanistan	44	Indonesia	7
Sudan	42	Nigeria	7
India	40	Rwanda	7
Colombia	37	Uganda	7
Pakistan	34	Egypt	5
Nepal	31	Israel	5
Chad	16	Lebanon	5
Algeria	11	Niger	5
Congo	11	Thailand	5
Russia	9	Iran	4
Philippines	8	Bangladesh	3

This information indicates how often each country is mentioned in the reports, and helps us to get some sense of what happened in the period covered by the reports.

The Noun Phrase Lexicon:

In order for a system to extract information from online documents, the system must know the syntactic structure of sentences. In general, a sentence (S) consists of a noun phrase (NP) and a verb phrase (VP). Some verb phrases do not require noun phrase objects but some do. A noun phrase can contain just a single noun (such as "civilian," "children," "hostage," "victims," etc.) or a noun with some modifiers (such as "civilian police officers", "deteriorating security situation", "guerrilla warfare", "five civilian vehicles", etc.). The following diagram



The Islamic State of Iraq (ISI) claimed responsibility

shows a very simple sentence structure:

Analysis of these syntactic structures helps CAINES to understand the semantic information contained in a sentence. In order to do sentence analysis automatically, the system should be able to determine the general linguistic units. We analyze documents using the part of speech tagger *Lingua::En::Tagger*, available at http://search.cpan.org/~acoburn/Lingua-EN-Tagger-0.16/Tagger.pm. *Lingua::En::Tagger* is a probability based, corpus-trained tagger that assigns Part-of-Speech (POS) tags to English text based on a lookup dictionary and a set of probability values. *Lingua::En::Tagger* has a method called, "get_words," that is very helpful in building our information extraction system. "Get_words" extracts all the possible maximal noun phrases from the input documents. For the portion of the text shown in Table 3 earlier, *Lingua::En::Tagger* produces the Part-of-Speech tags shown in Figure 2.

<in>On</in> <nnp>July</nnp> <cd>2</cd> <ppc>,</ppc> <nnp>Abdu</nnp> <nnp>Mohamad</nnp>
<jj>Sad</jj> <nnp>Ahmad</nnp> <nnp>Reheqa</nnp> <vbd>drove</vbd> <det>a</det>
<nn>suicide</nn> <nn>vehicle-borne</nn> <vbd>improvised</vbd> <jj>explosive</jj>
<nn>device</nn> <lrb>(</lrb> <nnp>SVBIED</nnp> <rrb>)</rrb> <in>into</in> <det>a</det>
<nn>group</nn> <in>of</in> <jj>western</jj> <nns>tourists</nns> <in>in</in> <nnp>Marib</nnp>
<ppc>,</ppc> <vbg>killing</vbg> <prp>him</prp> <cc>and</cc> <jj>several</jj> <nns>others</nns>
<pp>.</pp> <nnp>AQY</nnp> <vbd>claimed</vbd> <nn>responsibility</nn> <pp>.</pp>
<nnp>Three</nnp> <nns>days</nns> <rb>later</rb> <ppc>,</ppc> <nnp>U.S.-trained</nnp>
<nnp>Yemeni</nnp> <nn>security</nn> <nns>forces</nns> <vbd>killed</vbd> <det>the</det>
<jj>suspected</jj> <nn>leader</nn> <in>of</in> <det>the</det> <nnp>SVBIED</nnp>
<nn>bombing</nn> <ppc>,</ppc> <nnp>Ahmed</nnp> <nnp>Basyouni</nnp> <nnp>Dwedar</nnp>
<ppc>,</ppc> <det>an</det> <jj>Egyptian</jj> <vbd>wanted</vbd> <in>in</in> <nnp>Egypt</nnp>
<in>for</in> <prps>his</prps> <nns>ties</nns> <to>to</to> <det>the</det> <nnp>Muslim</nnp>
<nnp>Brotherhood</nnp> <pp>.</pp> <in>On</in> <nnp>August</nnp> <cd>8</cd> <cc>and</cc>
<cd>13</cd> <ppc>,</ppc> <nnp>Yemeni</nnp> <nn>security</nn> <nns>forces</nns>
<vbd>raided</vbd> <cd>two</cd> <nns>houses</nns> <ppc>,</ppc> <vbg>arresting</vbg>
<cd>17</cd> <cc>and</cc> <nn>killing</nn> <cd>four</cd> <jj>AQ-affiliated</jj>
<nns>suspects</nns> <in>while</in> <nn>suffering</nn> <nn>one</nn> <nn>casualty</nn>
<pp>.</pp>

FIGURE 2. Part-of-Speech tagging of the text in Table 3.

The noun phrases generated by *Lingua::En::Tagger* help the system to recognize the subjects and objects of sentences. Table 7 shows some noun phrases and the number of times they appear in the 2005 report.

TABLE 7. Noun Phrases and the Number of Times they appear in the 2005 Report

al- qa'ida in iraq	6
al-rafidayn	39
ansar al-sunnah	10
bomber detonated	17
bus station	4
command-initiated	5
department of state	7
headquarters	10
information	18
revolutionary armed forces of colombia	10
shia pilgrims	5
suicide bomber detonated	15
suicide bomber with a vehicle-borne	23

The Verb Phrase Lexicon:

To further improve the quality and accuracy of natural language processing, we must consider the meaning of the text at the core of each sentence. Main verbs describe an action (e.g. throw, hit) or state of being (e.g. feel, is) and are important because they assist us in building a system that extracts relevant information.

By conducting language analysis, the relations between the verb meaning and its arguments help the system better determine sentiment, emotions, and opinions [45]. In information extraction, being able to identify implied, but not stated assertions will benefit the accuracy of the applications. Repositories of semantic relations between verbs could benefit many natural language processing tasks [7]. The VerbOcean sytem [7], extracts semantic relations between verbs from the Web to detect similarity, strength, antonymy (semantic opposition), enablement, and happens-before relations. Some sample verb entries in our verb lexicon are: "attack", "blast", "bomb", "claim", "defend", "force", "kill", "protect", "resist", "strike", etc.

The Information Extraction Process

The Online Country Reports used in this study are domain specific. They have similar terms and patterns and thus can be used to create lexicons and extraction rules for use in CAINES. Sublanguage theory has assisted in analyzing documents with similar patterns in particular areas for years, such as in biomedicine [24, 41], weather forecasting [70], and others.

Sublanguage theory explains that texts common to a specific subject area will share common vocabulary, symbols, abbreviations, and even sentence construction [30, 33, 41, 42]. Thus, we prepare for extraction by applying knowledge engineering techniques to the reports to determine the format of the data presented and to detect patterns in the text. Knowledge engineering deals primarily with producing rules rather than training data. It is more precise but a major disadvantage is the dependence on the knowledge and skill of the engineer and the reliance on the test, re-test, and debug cycle [3].

Analysis of terms over nine out of twelve annual reports showed substantial overlap of specific and unique term usage in this domain: "forward-leaning", "governorates", "laundering/combating", and "tactical-level". Based on all we have learned by creating the lexicon discussed previously, we use the complete sublanguage lexicon to assist with analyzing the similar style, construction, and vocabulary of the online country reports. Abbreviations such as NCTC (National Counterterrorism Center) and AQ (al-Qa'ida) are also included.

This section discusses how our system extracts information from the reports. After we produce the KWIC index files for the reports, we are able to identify the key terms, N-Grams, and patterns in the reports, which helps us to generate the lexicons to be used in the information extraction process. Since these KWIC index files are in a MySQL RDBMS, it is easy to query the whole table and sort the data by the content in a given column. Table 8 shows the rows for which the word "killed" is in column 1.

Note however that, the key terms (words or phrases) may appear in other columns. The word "killing," in the following example appears in column 5 (shown in Table 9). This example also shows us how to identify the type of the attack (vehicle-borne improvised explosive device (VBIED)) and the result of the attack.

TABLE 8. A Portion of the KWIC index file that contains the word "killed" in column 1

killed	10	civilians.	No	Group	claimed	responsibility,	although	the
killed	101	people	and	wounded		1,200	others	in
killed	117	people	and	injured		200	others	in
killed	12	border	police	officers.		No	group	claimed responsibility.
killed	12	civilians	and	children		and	one	local Iraqi
killed	12	civilians	and	kidnapped		between	40	and 60
killed	40	civilians	by	unknown		means	and	wounded several

TABLE 9. A Portion of the KWIC index file that contains the word "killing" in column 5

vehicle-borne	improvised	explosive	device	(VBIED),	killing	10	civilians
vehicle-borne	improvised	explosive	device	(VBIED),	killing	12	students,
vehicle-borne	improvised	explosive	device	(VBIED),	killing	25	soldiers,
vehicle-borne	improvised	explosive	device	(VBIED),	killing	40	civilians,
vehicle-borne	improvised	explosive	device	(VBIED),	killing	46	Pakistani
vehicle-borne	improvised	explosive	device	(VBIED),	killing	eight	civilians,
vehicle-borne	improvised	explosive	device	(VBIED),	killing	the	Somali
vehicle-borne	improvised	explosive	devices	(VBIED),	killing	30	civilians,

By analyzing a great deal of data from many reports, we have been able to identify the key concepts and the patterns of how the reports were written. As discussed earlier, a sentence consists of a noun phrase and a verb phrase. We used the data generated in the KWIC index files to help us generate a set of noun and verb phrases that are useful in information extraction. To improve semantic interpretation, we also include more semantic relationships between terms (such as synonyms, hypernyms, etc.) as discussed by earlier authors [49, 50, 71] to expand the range of information that can be extracted.

Using the patterns found in the KWIC files with the general syntactic and semantic analyses; we have been able to generate rules for extraction. The following shows an example of a rule to identify the result of an attack:

for each row in the portion of the report from which we want to extract information
if the key word is a candidate denoting an attacking term (e.g., attack, damage, kill, wound, etc.)
then return the noun phrase immediately following the verb phrases and present it as the impact of the attack
end if
end for

Thus, using this simple rule, with the portion of the report containing the text shown in Figure 3, our CAINES system has been able to generate the output shown in Table 10.

A noun phrase that appears in front of the verb phrase in a sentence is the subject of that sentence. For example, the verb phrase "claimed responsibility" will have the noun phrase that appears in front of it being the actor. Thus, in the example shown in Table 11, the noun phrases "No group", "Pakistan (TTP)", and the "Taliban" are actors of the act "claimed responsibility".

> On 1 January 2008, late in the afternoon, in the Sab'ah Nisan district of Baghdad, Iraq, a suicide bomber detonated an improvised explosive device (IED) he was wearing among Sunni mourners in a house in the eastern Zayyunah neighborhood, killing at least 33 civilians and several members of Iraqi intelligence, wounding 38 civilians, and damaging the house and several vehicles. The Islamic State of Iraq (ISI) claimed responsibility.

Figure 3. Sample input to CAINES

TABLE 10. CAINES Output for Sample Text

Killed:	killing at least 33 civilians and several members of Iraqi intelligence
Wounded:	wounding 38 civilians
Damaged:	the house and several vehicles

TABLE 11. Sample Noun Phrases which are Actors of the Act "claimed responsibility"

No group	claimed responsibility,
Pakistan (TTP)	claimed responsibility.
The Taliban	claimed responsibility

Thus, in general, the rule for identifying the subject and object of an act is:

for each row in the portion of the report from which we want to extract information
if the key terms are a verb phrase
then return the noun phrase appearing immediately prior the verb phrase as an actor
and return the noun phrase immediately following the verb phrases as an object
end if
end for

Thus for the sentence "The Islamic State of Iraq (ISI) claimed responsibility", our system will return this information:

> **Terrorist Group Claiming Responsibility:** The Islamic State of Iraq (ISI)

Thus, with the portion of the input shown above, our system produces the extracted report shown in Table 12.

TABLE 12. CAINES Sample Output Report

Date:	1 January 2008
Place:	Sab'ah Nisan district of Baghdad, Iraq
Attacker:	a suicide bomber
Attack Type:	improvised explosive device (IED)
Killed:	killing at least 33 civilians and several members of Iraqi intelligence
Wounded:	wounding 38 civilians
Damage:	the house and several vehicles
Terrorist Group Claiming Responsibility:	The Islamic State of Iraq (ISI)

SYSTEM EVALUATION AND DISCUSSION

For the sample task described earlier in Figure 3 and Table 12, CAINES completed analysis in 3 seconds, compared to approximately 7 minutes for the same task for a human tagger. Anecdotally then, CAINES, once trained, executes significantly faster than a human tagger. Though net time savings are marginal or negative for small data sets (where training time is more significant than execution time), net time savings are significant for large data sets of hundreds or thousands of reports (where execution time is more significant than training time). However, time savings represent only one aspect of system evaluation. To assess CAINES accuracy and comprehensiveness, we undertook more rigorous and extensive tests, as follows.

Performance measures for IE systems were developed and refined at MUC's. Precision and recall performance metrics were developed for MUC-3 and MUC-4 to set a standard performance measure for IR and IE systems.

Precision measures reliability and accuracy by determining what percentage of the information extracted is correct [72]. Precision is calculated by dividing the total number of correctly extracted items by the total number of extracted items.

$$\text{Precision} = \frac{\text{Total correctly extracted}}{\text{Total extracted}}$$

For example, suppose the extraction template has 10 slots, and the domain experts are able to find answers to fill all 10 slots. If the system finds 10 answers for the 10 slots, but only 6 are correctly filled, then the precision rate is 6/10 = 60%.

Recall measures what percentage of the available correct information is extracted, thus measuring the ability of the system to extract relevant information [1]. Recall is the number of correct answers produced divided by the total possible correct answers.

$$\text{Recall} = \frac{\text{Total actually correctly extracted}}{\text{Total possible to be correctly extracted}}$$

For example, if system extracts 8 correct slot values and the total possible correct slot values is 10, the recall of that system is 80% (8/10).

F-measure, a combined weighted measure of precision and recall, was used at MUC-5. Per [46], a higher F-measure indicates greater performance. The F-measure combines recall and precision into a single measure by using the harmonic mean of precision and recall [3, 72]:

F-measure = 2 *(recall * precision) / (recall + precision)

By the mid-1990s TIPSTER and MUC systems showed average recall performance of 40%, with precision performance somewhat better at 50%. Some simple systems can reach performance levels in the 90% range [10]. Various sub tasks of IE had attained the performance level in the range of 60-80% by late 1990s [3].

Evaluation of CAINES in IE Terrorism Information Extraction

We evaluated CAINES by comparing the output of the system and the answers that the authors found through manual reading and extraction of relevant information from the same country reports. We ran CAINES using three country reports (2005 – 292 pages, 2007 – 312 pages, and 2009- 299 pages). We measured the system using recall, precision, and F-measure values as follows:

$$\textit{Precision} = \frac{\text{The number of items correctly tagged by CAINES}}{\text{The number of items tagged by CAINES}}$$

CAINES' Precision = 92%

$$\textit{Recall} = \frac{\text{The number of items correctly tagged by CAINES}}{\text{The number of possible items that the authors tagged}}$$

CAINES' Recall = 89%

F-measure = 2(R*P)/(R + P)
CAINES' F-measure = 90.48 %

CONCLUSION, DISCUSSION, AND FUTURE WORK

While information extraction has been previously attempted for news articles about terrorism in Latin America [56, 57, 58], this study, to our knowledge, is the first to employ information extraction techniques on the United States Department of State's Country Reports on Terrorism. Unlike many systems that use statistical analysis techniques, our system, CAINES, is based on a knowledge engineering approach. It relies on lexicons of various categories of terms, as well as sublanguage analysis, to perform syntactic and semantic analysis. This allows CAINES to use the syntactic and semantic structure of language in analyzing textual reports, rather than just using statistics. We demonstrated that CAINES is able to rapidly and efficiently extract relevant and precise information from these reports. CAINES can extract relevant and precise national intelligence information in much less time than manually reading reports on terrorism.

Further work could attempt to generalize CAINES to other information sources, such as online news reports from the popular press, or facebook or twitter feeds that may be of interest to intelligence services. The ability to rapidly extract structured information from unstructured narratives transforms such textual reports from relatively opaque data artifacts, to pliable information assets, that become more amenable to various descriptive and predictive analytics tools. These analytics tools have attracted substantial attention in the defense intelligence community [48]. For instance, following successful information extraction from unstructured sources into structured form, ad hoc descriptive managerial analytics reports showing "incidents by time of day", "incidents by day of week", "incidents by place", "incidents by weapon used", "deaths per incident", and so forth, could be rapidly assembled with satisfactory accuracy, from textual reports. Furthermore, data mining techniques – used for predictive analytics – such as association rules, decision trees, logistic regression, and neural nets [21] could allow military strategists to predict likely weapon types by terrorist group, likely weapon types by location, likely times of day or day of week by terrorist group, or other patterns that may assist with resourcing decisions and mitigation tactics. In future work, we hope to build descriptive and predictive analytics services as a value-added layer on top of our information extraction platform, and to determine whether these can provide valuable intelligence from textual narrative sources on the web with sufficient speed and accuracy.

REFERENCES

[1] Adams, K. 2001. The Web as a database: New extraction technologies and content management. Online (25:2), 2001, 27–32.

[2] Andersen, P. M., Hayes, P. J., Heuttner, A. K., Schmandt, L. M., and Nirenberg, I. B. Automatic extraction. In Proceedings of the Conference of the Association for Artificial Intelligence, Philadelphia, PA, 1986, 1089–1093.

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

[4] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. "Open information extraction for the web," Proceedings of the 20th international joint conference on Artical intelligence, Hyderabad, India: 2007, 2670-2676.

[5] Chang, C.H., Kayed, M., Girgis, R., and Shaalan, K. F. "A survey of web information extraction systems," Knowledge and Data Engineering, IEEE Transactions (18:10), 2006, 1411-1428.

[6] Chaudhuri, S., and Dayal, U. "An overview of data warehousing and OLAP technology," ACM SIGMOD Record (26:1), 1997, 65-74.

[7] Chklovski T. and Pantel P. "VERBOCEAN: Mining the Web for Fine-Grained Semantic Verb Relations," Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04). Barcelona, Spain, 2004, 33-40.

[8] Ciravegna, D. "Adaptive information extraction from text by rule induction and generalisation," Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2, 2001, 1251-1256.

[9] Conlon, S., Hale, J., Lukose, S., and Strong, J. "Information extraction agents for service-oriented architecture using web service systems: a framework," Journal of Computer Information Systems (48:3), 2008, 74-83.

[10] Cowie, J., and Lehnert, W. "Information extraction,"

Communications of the ACM (39:1), 1996, 80-91.
[11] Cowie, J., Wakao, T., Guthrie, L., Jin, W., Pustejovsky, J., and Waterman, S. "The Diderot information extraction system," The First Conference of the Pacific Association for Computational Linguistics, Vancouver, British Columbia, Canada: 1993, 5-14.
[12] Cowie, J., and Wilks, Y. "Information Extraction," in Handbook of Natural Language Processing. , R. Dale, H. Moisl and H. Somers (eds.), Marcel Dekke: New York, 2000, 241-260.
[13] Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., and Slattery, S. "Learning to construct knowledge bases from the World Wide Web," Artificial Intelligence (118:1-2), 2000, 69-113.
[14] Crescenzi, V., Mecca, G., and Merialdo, P. "Roadrunner: Towards automatic data extraction from large web sites," Proceedings Of The International Conference On Very Large Data Bases, 2001, 109-118.
[15] Cunningham, H. "Information extraction, automatic," Encyclopedia of Language and Linguistics), 2005, 665-677.
[16] Das, D., Schneider, N., Chen, D., and Smith, N. A. "SEMAFOR 1.0: A probabilistic frame-semantic parser," Technical Report CMU-LTI-10-001, Carnegie Mellon University, 2010.
[17] Dingli, A., Ciravegna, F., and Wilks, Y. Year. "Automatic semantic annotation using unsupervised information extraction and integration," Proceedings of SemAnnot 2003 Workshop, 2003.
[18] Eikvil, L. "Information Extraction From World Wide Web - A Survey", Technical Report 945, Norweigan Computing Center, 1999.
[19] Etzioni, O. "The World-Wide Web: quagmire or gold mine?" Communications of the ACM (39:11), 1996, 65-68.
[20] Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. "Open information extraction from the web," Communications of the ACM (51:12), 2008, 68-74.
[21] Fayyad, U., Shapiro, G.P., and Smyth, P. "From data mining to knowledge discovery in databases," AI magazine (17:3), 1996, 37-54.
[22] Florescu, D., Levy, A., and Mendelzon, A. "Database techniques for the World-Wide Web: A survey," SIGMOD record (27:3), 1998, 59-74.
[23] Freitag, D. "Machine learning for information extraction in informal domains," Machine learning (39:2-3), 2000, 169-202.
[24] Friedman, C., Kraa, P., & Rzhetskya, A. "Two Biomedical Sublanguages: a Description Based on the Theories of Zellig Harris," Journal of Biomedical Informatics (35:4), 2002, 222-235.
[25] Gaizauskas, R., and Wilks, Y. "Information extraction: Beyond document retrieval," Journal of Documentation (54:1), 1998, 70-105.
[26] Gerdes J. "EDGAR-Analyzer: automating the analysis of corporate data contained in the SEC's EDGAR database," Decision Support Systems (35:1), 2003, 7-29.
[27] Gildea, D., and Jurafsky, D. "Automatic labeling of semantic roles," Computational Linguistics (28:3), 2002, 245-288.
[28] Grant, G. H., and Conlon, S. J. "EDGAR extraction system: An automated approach to analyze employee stock option disclosures," Journal of Information Systems (20:2), 2006, 119-142.
[29] Grishman, R. "Information extraction: Techniques and challenges," in Information Extraction A Multidisciplinary Approach to an Emerging Information Technology, Springer, 1997, 10-27.
[30] Grishman, R. & Kittredge, R. I. (Eds.). Analyzing Language in Restricted Domains: Sublanguage Description & Processing: Lawrence Erlbaum Associates, 1986.
[31] Grishman, R., and Sundheim, B. "Message Understanding Conference-6: A Brief History." In COLING-96, Copenhagen, Denmark, 1996, 466-471.
[32] Gupta, V., and Lehal, G. S. "A survey of text mining techniques and applications," Journal of Emerging Technologies in Web Intelligence (1:1), 2009, 60-76.
[33] Harris, Z. A theory of language and information: a mathematical approach. Clarendon Press, Oxford, 1991.
[34] Hobbs, J., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M., and Tyson M. "FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text." Finite-State Language Processing, 1997, 383-406.
[35] Isa M. and Vossen P. "A verb lexicon model for deep sentiment analysis and opinion mining applications," Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT, Portland, Oregon, 24 June, 2011, 10–18.
[36] Jacobs, P. and Rau, L. "SCISOR: Extracting information from on-line news." Communications of the ACM, (33:11), 1990, 88-97.
[37] Kushmerick, N. "Wrapper induction: Efficiency and expressiveness," Artificial Intelligence (118:1), 2000, 15-68.
[38] Laender, A. H., Ribeiro-Neto, B. A., da Silva, A. S., and Teixeira, J. S. "A brief survey of web data extraction tools," ACM SIGMOD Record (31:2), 2002, 84-93.
[39] Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E., and Soderland, S. Year. "University of Massachusetts: MUC-4 test results and analysis," Proceedings of the 4th conference on Message understanding, Association for Computational Linguistics, 1992, 151-158.
[40] Leinemann, Christoph, Frank Schlottmann, Detlef Seese, and Thomas Stuempert. "Automatic extraction and analysis of financial data from the EDGAR database," South African Journal of Information Management (online), (3:2), 2001.
[41] Liddy, E. D., Jorgensen, C. L., Sibert, E. E. & Yu, E. S. "A Sublanguage Approach to Natural Language Processing for an Expert System," Information Processing & Management (29:5), 1993, 633-645.
[42] Liddy, E. D., Symonenko, S., Rowe, S. Sublanguage analysis applied to trouble tickets. Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference, 2006, 752-757.
[43] Luhn, H. P. "Keyword-in-context index for technical literature (KWIC index)," American Documentation 11, 1960, 288-295.
[44] Lytinen, S.L., and Gershman A. "ATRANS Automatic Processing of Money Transfer Messages," AAAI, 1986, 1089-1095
[45] Maks, I., and Vossen P. "A verb lexicon model for deep sentiment analysis and opinion mining applications," In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Association for Computational Linguistics, Stroudsburg, PA: 2011, 10-18.
[46] Manning, C., & Schutze, H. Foundations of statistical natural language processing (5th ed.). Cambridge, MA: MIT Press, 2002.
[47] McCallum, A. "Information extraction: Distilling structured data from unstructured text," Queue (3:9), 2005, 48-57.
[48] McCue, C. "Data Mining and Predictive Analytics: Battlespace Awareness for the War on Terrorism," Defense Intelligence Journal (13:1), 2005, 47-63.
[49] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. "Introduction to WordNet: An on-line lexical database," International Journal of Lexicography (3:4),

1990, 235-244.
[50] Miller, G. A., & Fellbaum, C. "Semantic networks of English," In B. Levin & S. Pinker (Eds.), Lexical and conceptual semantics (41:1-3), 1991, 197-229.
[51] Moens, M.F. Information extraction: algorithms and prospects in a retrieval context, Springer, 2006.
[52] Muslea, I. Year. "Extraction patterns for information extraction tasks: A survey," The AAAI-99 Workshop on Machine Learning for Information Extraction, Orlando Florida: 1999.
[53] Nahm, U. Y., and Mooney, R. J. Text mining with information extraction, AAAI Technical Report SS-02-06, 2002.
[54] Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A., and Goranov, M. Year. "Towards semantic web information extraction," Human Language Technologies Workshop at the 2nd International Semantic Web Conference (ISWC2003), Sanibel Island, Florida: 2003.
[55] Punyakanok, V., Roth, D., and Yih, W. "The importance of syntactic parsing and inference in semantic role labeling," Computational Linguistics (34:2), 2008, 257-287.
[56] Riloff, E. M. "Automatically constructing a dictionary for information extraction tasks," Proceedings of the National Conference on Artificial Intelligence, John Wiley & Sons Ltd, 1993, 811-816.
[57] Riloff, E. M. "Little words can make a big difference for text classification," Proceedings Of The 18th Annual International ACM SIGIR Conference On Research And Development In Information Retrieval, ACM, Seattle, WA, 1995, 130-136.
[58] Riloff, E. M. "Automatically generating extraction patterns from untagged text," Proceedings Of The National Conference On Artificial Intelligence, Portland, Oregon: 1996, 1044-1049.
[59] Riloff, E. M. "An empirical study of automated dictionary construction for information extraction in three domains," Artificial Intelligence (85:1), 1996, 101-134.
[60] Riloff, E., and Phillips, W. "An introduction to the Sundance and Autoslog systems," Technical Report UUCS-04-015, School of Computing, University of Utah, 2004.
[61] Sager, N., Friedman, C., and Lyman, M., Medical Language Processing: Computer Management of Narrative Data. Addison Wesley, 1987.
[62] Saggion, H., Funk, A., Maynard, D., and Bontcheva, K. "Ontology-based information extraction for business intelligence," in The Semantic Web, Springer, 2007, 843-856.
[63] Sarawagi, S. "Information extraction," Foundations And Trends In Databases (1:3), 2008, 261-377.
[64] Sharples, M., Hogg, D., Hutchinson, C., Torrance, S., & Young, D. Computers And Thought: A Practical Introduction To Artificial Intelligence, The MIT Press, 1989.
[65] Sheikh, M., and Conlon, S. "Use of a Fast Information Extraction Method as a Decision Support Tool," Journal of International Technology and Information Management (19:4), 2010.
[66] Sheikh M. and Conlon S., "A Rule-Based System to Extract Financial Information," Journal of Computer Information Systems (52:4), 2012, 10-19.
[67] Simmons L.L., Conlon S., Mukhopadhyay S., and Yang J. "A Computer Aided Content Analysis of Online Reviews," Journal of Computer Information Systems (52:1), 2011, 43-55.
[68] Simmons L.L., and Conlon S. "Extraction Of Financial Information From Online Business Reports," Database: The Journal of the ACM Special Interest Group on Management Information Systems (SIGMIS) (44:3), 2013, 34-48.
[69] Soderland, S. "Learning information extraction rules for semi-structured and free text," Machine learning (34:1-3), 1999, 233-272.
[70] Somers, H. (ed.), Computers and Translation: A translator's guide, John Benjamins Publishing Company, 2003, 217-219.
[71] Sowa, J. F. Knowledge representation: Logical, philosophical, and computational foundations. Pacific Grove, CA: Brooks/Cole, 2000
[72] Van Rijsbergen, C. J. Information retrieval (2nd ed.). London: Butterworths, 1979.
[73] Wilks, Y. Information Extraction as a core language technology. In M-T. Pazienza (ed.), Information Extraction, Springer, Berlin, 1997.