Rule-based Information Extraction from Disease Outbreak Reports

Wafa N. Alshowaib *walshowib@kacst.edu.sa*
National Center for Computation Technology and Applied Mathematics
KACST
Riyadh, 11442, Saudi Arabia

Abstract

Information extraction (IE) systems serve as the front end and core stage in different natural language programming tasks. As IE has proved its efficiency in domain-specific tasks, this project focused on one domain: disease outbreak reports. Several reports from the World Health Organization were carefully examined to formulate the extraction tasks: named-entities, such as disease name, date and location; the location of the reporting authority; and the outbreak incident. Extraction rules were then designed, based on a study of the textual expressions and elements found in the text that appeared before and after the target text.

The experiment resulted in very high performance scores for all the tasks in general. The training corpora and the testing corpora were tested separately. The system performed with higher accuracy with entities and events extraction than with relationship extraction.

It can be concluded that the rule-based approach has been proven capable of delivering reliable IE, with extremely high accuracy and coverage results. However, this approach requires an extensive, time-consuming, manual study of word classes and phrases.

Keywords: Information Extraction, Disease Outbreak, Rule-based, NLP.

1. INTRODUCTION

With the tremendous amount of data that accumulate on the web every second, the urge for automatic technologies that read, analyze, classify and populate data has evolved. Humans cannot read and memorize a megabyte of data on a daily basis. This has resulted in opportunities for historical, archival information to be lost or discarded. Information that may currently seem to contain no value may hold valuable information for future needs. Information also runs the risk of being overlooked or missed because it was not presented in a specific manner or was contained with additional misleading data.

Lost opportunities and limited human abilities have spurred researchers to explore and create strategies to manage this text 'wilderness'. In the last decades, researchers have mainly worked in natural language techniques. Since human language is difficult and follows different writing styles, the Natural Language Processing (NLP) technologies cannot be classified under one domain only. Different stages of processing comprise the NLP field, and each stage is a unique science and field of research. IE systems serve as the front-end and core stage in different NLP techniques.

In the literature, different researchers give different descriptions for the term 'Information Extraction' (IE). One of the oldest definitions was proposed by Cowie and Lehnert, who define it as any process that extracts relevant information from a given text then pieces together the extracted information in a coherent structure [1]. De Sitter sees that IE can take a different definition according to the purpose of the system: One best per approach: the information system

is a system for filling a template structure; All occurrences approach: the IE Information system is to find every occurrence of a certain item in a document [2].

However, De Sitter's definition lacks the part about recognizing relationships and facts. Moens suggests a very comprehensive definition:

"Information extraction is the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in unstructured data sources, such as natural language text, providing additional aids to access and interpret the unstructured data by information systems." [3]

It seems that in recent IE manuscripts, researchers partially agree with similar descriptions. Ling described an IE system as a problem of distilling relational data from unstructured texts [4].

Acharya and Parija suggested another definition, which is to reduce the size of text to a tabular form by identifying only subsets of instances of a specific class of relationships or events from a natural language document, and the extraction of the arguments related to the event or relationship [5].

Before continuing with discussion in this paper, it seems essential to view the definition that has been adopted for this project. We agree with Moens' definition that additional aids are needed to find primary data [3], and also with De Sitter from the aspect that IE can handle more than one definition depending on the aim of the system [2].

The definition that seems most comprehensive for this project is that IE is the process of extracting predefined entities, identifying the relationships between those entities from natural texts into accessible formats that can be used later in further applications and, with the help of evidence, can be deduced from particular words from the text or from the context.

2. INFORMATION EXTRACTION TASKS

The prime goal of IE has been divided into several tasks. The tasks are of increasing difficulty, starting from identifying names in natural texts then moving into finding relationships and events.

2.1. Named Entity Recognition

The term *named entity recognition* (NER) was first introduced in Message Understanding conferences MUCs [6]. A key element of any extraction system is to identify the occurrence of specific entities to be extracted. It is the simplest and most reliable IE subtask [7]. Entities typically are noun elements that can be found within text, and they usually consist of one to a few words. In early work in the field, more specifically at the beginning of the the Message understanding conferences (MUC) and Automatic Content Extraction (ACE) competitions, the most common entities were named entities, such as names of persons, locations, companies and organizations, numeric expressions, e.g. $1 million, and absolute temporal terms, e.g. September 2001. Now, named entities have been expanded to include other generic names, such as names of diseases, proteins, article titles and journals. More than 100 entity types have been introduced in the ACE competition for named entity and relationship extraction from natural language documents [8].

The NER task not only focuses on detecting names, but it can also include descriptive properties from the text about the extracted entities. For instance, in the case of person names, it can extract the title, age, nationality, gender, position and any other related attributes [9].

There is now a wide range of systems designed for NER, such as the Stanford Named Entity Recognizer[1]. Regarding the performance of these subsystems, the accuracy reached 95 per cent. However, this accuracy only applies for domain-dependent systems. To use the system for extracting entities from other types, changes must be applied [7].

2.2. Relationship Extraction

Another task of the IE system is to identify the connecting properties of entities. This can be done by annotating relationships that are usually defined between two or more entities. An example of this is 'is an employee of', which describe the relationship between an employee and a company; 'is caused by' is a relationship between an illness and a virus [8]. Although the number of relations between entities that may be of interest can generally be unlimited, in IE, they are fixed and previously defined, and this is considered part of achieving a well-specified task [10]. The extraction of relations differs completely from entity extraction. This is because entities are found in the text as sequences of annotated words, whereas associations are expressed between two separate snippets of data representing the target entities [8].

2.3. Event Extraction

Extracting events in unstructured texts refers to identifying detailed information about entities. These tasks require the extraction of several named entities and the relationships between them. Mainly, events can be detected by knowing who did what, when, for whom and where [11].

3. IE SYSTEM PERFORMANCE

One of the main outputs of the MUC series that took place between 1987 and 1997 was defining the evaluation standards for IE systems. For instance, defining quantitative metrics, such as precision and recall. In total there were seven conferences.

Measuring the overall performance of IE systems is an aggregated process and it depends on multiple factors. The most important factors are (i) the level of the logical structure to be detected, such as named entities, relationships, events and the co-references, (ii) the type of system input, i.e. newspapers articles, corporate reports, database tuples, short text messages from social media or sent by mobile phones, (iii) the focus of the domain, i.e. political, medical, financial, natural disasters, (iv) the language of the input texts, i.e., English, or a more morphologically sophisticated language, such as Arabic [10].

The relative complexity of assessing the performance of an IE system can be managed by noting the scores obtained by the system in all the MUC events. In the last event, the MUC-7, in which the domain in focus was aircraft accidents from English newspaper articles, the system, which achieved the highest overall score, obtained a different score for each subsystem. Scores for both recall and precision are presented in table 1 [10]. These figures provide a glimpse of what to expect from an IE system in which the best performance is achieved in NER and the lowest scores indicate the most difficult task of event extraction.

[1] Stanford Named Entity Recognizer website: http://nlp.stanford.edu/software/CRF-NER.shtml [Last Accessed: 12 June 2014]

Task	Recall score	Precision score
NER	95	95
Relationships	70	85
Events	50	70
Co-reference	80	60

TABLE 1: Top Scoring in MUC-7.

4. DOMAIN SELECTION

It is necessary to complement the clinical-based reporting systems by enriching their databases with information extracted from disease outbreak reports. Information related to diseases outbreaks is often written as free text, and, therefore, difficult to use in computerized systems. Confining such information in this rigid format makes the process of accessing it rapidly very difficult.

According to the World Health Organization (WHO), analyzing the information from disease epidemic reports can be used for:

- The identification of disease clusters and patterns;
- Facilitating the tracking and following up with the spread of a disease outbreak;
- Estimating the potential increase in the number of infected people in a further spread;
- Providing an early warning in the case of an increase in the number of incidents;
- Help in strategic decision-making as to whether control measures are working effectively.

In addition to these factors, it has been found throughout history that the number of information systems that were designed to extract information from disease outbreak reports is very few and limited, the only system that was designed for disease outbreaks is Proteus-BIO in 2002 [12]. All these are the motivating factors for choosing to study this domain.

The intention of the work proposed in this project is to be able to extract information about disease outbreaks from different natural texts. There are a number of news websites that produce reports about disease outbreaks. Some of these reports are annual reports, where one report presents a summary of all disease epidemics that have been recorded in one year; for example, reports found in the Centre for Disease Control and Prevention website[2] are all historical reports related to food-borne diseases. However, for the sake of the project, the aim is to analyze texts from different formats, where one particular disease is discussed in many reports. Therefore, a decision has been taken to mainly analyze news reports from WHO. The WHO website[3] represents an ideal source that contains archives of news classified in different categories, either by country, year or by disease.

In almost every case, the authors of these news stories are reporting the same information; however, from report to report, the style of writing is slightly different. This provides an opportunity of being exposed to a variety of writing styles in reporting disease data.

[2] http://www.cdc.gov/foodsafety/fdoss/index.html [Last Accessed: 12 June 2014]

[3] http://www.who.int/csr/don/en/ [Last Accessed: 12 June 2014]

The following are examples of news taken from the WHO:

• **Example 1:** A sample of reporting incidents in Brazil diagnosed with dengue haemorrhagic fever:
"10 April 2008 - As of 28 March, 2008, the Brazilian health authorities have reported a national total of 120 570 cases of dengue including 647 dengue haemorrhagic fever (DHF) cases, with 48 deaths.

On 2 April 2008, the State of Rio de Janeiro reported 57 010 cases of dengue fever (DF) including 67 confirmed deaths and 58 deaths currently under investigation. Rio de Janeiro, where DEN-3 has been the predominant circulating serotype for the past 5 years since the major DEN-3 epidemic in 2002, is now experiencing the renewed circulation of DEN-2. This has led to an increase in severe dengue cases in children and about 50% of the deaths, so far, have been children of 0-13 years of age.

The Ministry of Health (MoH) is working closely with the Rio de Janeiro branch of the Centro de Informações Estratégicas em Vigilância em Saúde (CIEVS) to implement the required control measures and identify priority areas for intervention. The MoH has already mobilized health professionals to the federal hospitals of Rio de Janeiro to support patient management activities, including clinical case management and laboratory diagnosis.

Additionally public health and emergency services professionals have been recruited to assist community-based interventions. Vector control activities were implemented throughout the State and especially in the Municipality of Rio. The Fire Department, military, and health inspectors of Funasa (Fundacao Nacional de Saude, MoH) are assisting in these activities." [4]

• **Example 2:** A sample of reporting incidents in Turkey diagnosed with Avian influenza:
"*30 January 2006 - A WHO collaborating laboratory in the United Kingdom has now confirmed 12 of the 21 cases of H5N1 avian influenza previously announced by the Turkish Ministry of Health. All four fatalities are among the 12 confirmed cases. Samples from the remaining 9 patients, confirmed as H5 positive in the Ankara laboratory, are undergoing further joint investigation by the Ankara and UK laboratories. Testing for H5N1 infection is technically challenging, particularly under the conditions of an outbreak where large numbers of samples are submitted for testing and rapid results are needed to guide clinical decisions. Additional testing in a WHO collaborating laboratory may produce inconclusive or only weakly positive results. In such cases, clinical data about the patient are used to make a final assessment." 5*

After reviewing 25 outbreak reports from the WHO website, it can be said that most of them follow a general scheme. Reports chosen for this project will range from 100 to 300 words in length, because those of over 300 words usually contain additional information, such as recommendations or medical treatments, which are out of the scope of this study.

4.1. Preprocessing

All the documents on the WHO website start with a date string indicating the data of publication. Some elements are common to all texts, such as information about the number of people affected by an outbreak, the name of the disease, and the location where it is spreading. The structure usually consists of the following points:

- The first sentence ,after the title, contains a date string featuring the date of publication on the website, always presented in the same format, e.g. 2 April 2011.

[4] http://www.who.int/csr/don/2008_04_10/en/index.html [Last Accessed: 12 June 2014]
[5] http://www.who.int/csr/don/2006_01_30/en/index.html [Last Accessed: 12 June 2014]

• Some of the reports contain another date, which is the announcement date; this is always given after the first date in the text, in the second sentence in most cases.

• In most reports the disease is reported by a health agency in a country, e.g. “The Brazilian health autorities “. This can be very useful piece of information, since it has been noticed that in some reports the name of country is not mentioned, and the name of the national health agency is enough to indicate the location.

• Disease names are not capitalized but they are sometimes accompanied with indicating words such as fever, outbreak, influenza etc. Some of the disease names are combination of characters and numbers like H5N1-influenza.

• The report identifies the number of suspected and confirmed disease cases.

• The total number of people affected by the disease since it was initially discovered to the date of announcement is sometimes reported.

• Infected cases are reported individually for one state, from the text: the State of Rio de Janeiro reported 57 010 cases of dengue fever (DF) including 67 confirmed deaths and 58 deaths currently under investigation.

• A pattern of “health authority of a Country *reported* victims” is very common, such as “the Brazilian health authorities have reported a national total of 120 570”. In some cases the word reported is replaced by synonyms such as “announced”. Also, this pattern can be found in other documents in other order “victims reported by health agency of country”.

4.2. Entities, Relationships and Events Identification

Information about a particular incident must be drawn from several constituent elements within the text: the publication date, announcement date, disease name, country of the outbreak, specific location (cities, states), number of infected people, status of victims (sick, dead). The organization of this information often differs from one report to another.

To make event extraction easier, we need to distinguish between relationships and events. Relationship extraction will rely on identifying a single piece of information, such as the nationality of the reporting authority, while the event will be the outbreak incident itself (number of people infected by the disease in the country).

In sum, the extraction task will involve detecting the following information elements:

Entities

Publication date
Announcement date
Disease name
Disease code
Country
Locations of the outbreak(cities, villages,...)

Relationships

Nationality of the reporting authority.

Events

Number of cases and deaths of an outbreak.
Total number of affected cases.

Entity	Position in the text
Report date	Actual string from document
Disease name	Clue words: fever, outbreak, syndrome, influenza, fever
Health agency name	Actual string from document, clues: ministry, agency
Country	Actual string from document, or computed from the agency name
Location	States, cities, town
Number of victims	Numeric value of cases mentioned in the text, or computed by counting number of cases in different locations

TABLE 2: Named Entities In Outbreak Reports for IE.

5. EXTRACTION ENGINE

The CAFETIERE system is a rule-based system for the detection and extraction of basic semantic elements. CAFETIERE is an abbreviated term for Conceptual Annotation for Facts, Events, Terms, Individual Entities and RElations. It is an information extraction system developed by the National Center for Text Mining at the University of Manchester. The engine incorporates a knowledge engineering approach for extracting tasks. In this project, CAFETIERE is the extraction engine to be used.

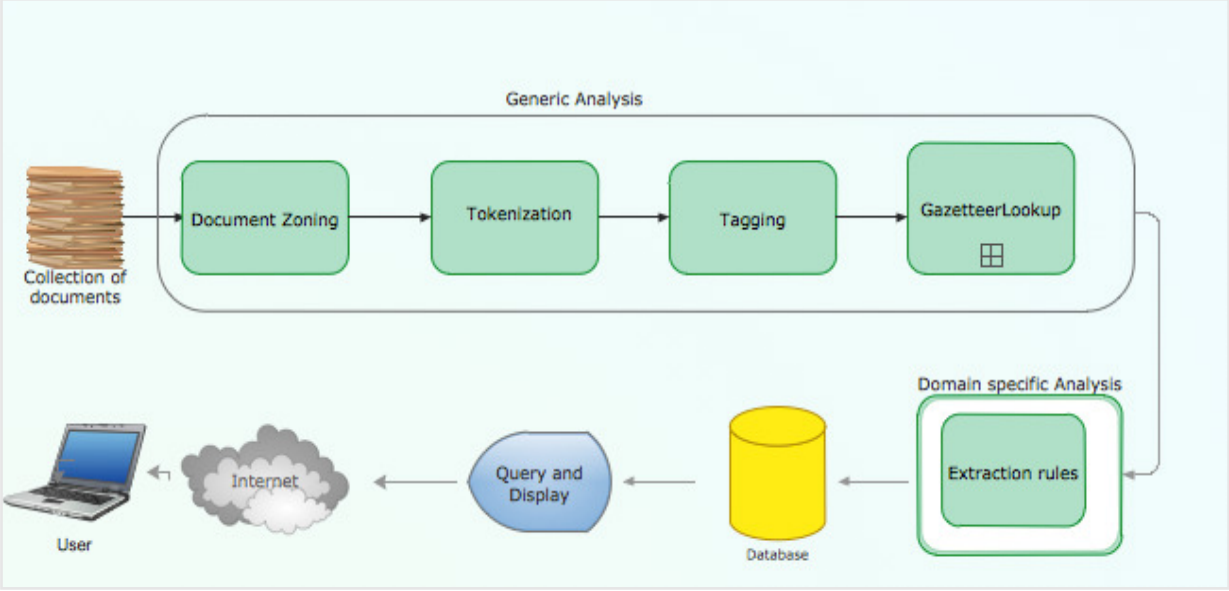


FIGURE 1: CAFETIERE Overall Analysis and Query Model.

5.1. Notation of the Rules

Rules in cafeteria system are written in high-level programming language [13]. The general rule formalism in CAFETIERE is [13]:

Y => A \ X / B

Where Y represents the phrase to be extracted and X are the elements that are part of the phrase. A represents the part of the context that may appear immediately before X in the text, and B represents the part of context that may appear immediately after X, both A and B are optional (which means may be null). Many rules lack the presence of A or B or both, therefore rules may have one of the following form [13]:

Y => \ X / B
Y => A \ X /
Y => \ X /

Rules are context sensitive, that means the constituents present in the text before and after the phrase should be reflect on the right hand-side (***A)*** and on the left hand-side (***B)*** of the rule. The rule must have at least one constituent (X) and can be more if required.

These rules define phrases (*Y*) and their constituents (*A, X, B*) as pairs of features and their corresponding values, for example:

Y	**A**	**X**	**B**
[syn=np, sem=date] =>		\ [syn=CD], [sem=temporal/ interval/month], [syn=CD]	

Where this represents a context free rule where both A and B are null, the phrase and the constituent part are written in as a sequence of features and values enclosed by in square brackets [Feature Operator Value]. If there is more than one feature, then they are separated by commas, as can be seen in the feature bundle for Y . Following is brief description for each of them:

• **Feature**: Denote the attribute of the phrase to be extracted. The most commonly used features are syn, sem, and orth, where syn is syntactic, sem is semantic and orth is orthography. Features are written as sequence of atomic symbols.
For example, some of the values (and their meaning) may be assigned to the feature syn listed in table 3 [13].

Tag	**Category**	**Example**
CD	Cardinal number	4 , four
NNP	Proper noun	London
NN	Common noun	girl, boy
JJ	Adjective	happy, sad

TABLE 3: Examples of the values that can be assigned to syn feature.

Although these features are built into the system, there is no restriction on the name of the features on the left-hand side of the rule.

• **Operator**: Denote the function applied to the attribute and the predicated value, operators that can be used in the system are (>, >=, <, <=, , =, !=, ~) all have the same usual meaning, the tilde operator matches a text unit with a pattern.

• **Value:** expresses a literal value that may be strings or numbers or combination of both, they may be quoted or unquoted.

5.2. Gazetteer

In addition to the system built-in gazetteer, users can upload their own gazetteers to look up words and expressions from the domain that they are focusing on. The system recognizes a plain text files with the extension .gaz as a gazetteer file, then it will added to the existed file (which is a relational database table) [13]. The look up mechanism of the gazetteer works by identifying all the strings that happen to be in the gazetteer by loading the relevant data from the lexical base of the system before applying the rules. This process may consume time because all tokens are looked up to see if there is any information about them in the gazetteer [14].

6. DESIGN AND IMPLEMENTATION

The process of designing the extraction rules is based on studying the textual expressions and elements found in the text, so for every entity, relationship and event, a similar approach has been followed. The following are the factors that influenced the design of the majority of the rules:

- Every textual element is recognized and captured using linguistic features (e.g. syntactic, semantic, orthography). For example, to extract a token of type number such as '45', the rule should contain the syntactic feature 'syn=CD'. (CD refers to Cardinal Number).

- For each extraction task, the span of text that appears before and after the target text is collected and studied to find common patterns that may help in identifying the correct element. This task of studying the context surrounding the element is the heart of this work as it is the only way to avoid false matches.

- Patterns can be very simple, such as 'prepositional phrase + noun', or very complex, such as the patterns used to look for outbreak events when the pattern is a whole sentence: 'verbs + prepositional phrases + nouns + punctuations'. Hence, not all the constituents mentioned in the pattern will be extracted - only the required ones.

- Rule order is very important. If there are two elements to be extracted and the first element depends on the existence of the second element in the sentence, then the first element should be extracted before the second one. This is because when an element is recognized by a rule, it will be hidden from the rest of the rules; therefore, each element is only extracted once.

Although the project uses rule-based systems, for some extraction tasks there was an essential need for additional entries to the system gazetteer. Therefore, one of the initial steps was to collect domain-specific vocabularies and then add them under the appropriate semantic class or create new semantic classes if needed. Not only have the domain terminologies been added to the gazetteer, but some commonly used verbs and nouns have also been collected, added and categorized.

6.1. Entity Extraction

6.1.1. Publishing Date

More than one date can be found for any randomly chosen outbreak report, each of which may refer to something different (such as the date of the first suspected ill person). To resolve this issue, we found that the publishing dates are usually mentioned in the first or second sentence; therefore, the extraction task here was restricted to only those locations.

6.1.2. Announcement Date

Expressions such as 'As of 6 July 2002, the ministry of health has reported . . .' and 'On 4 March, the Gabonese Ministry of Public Health reported . . .' are used to report the news; therefore, the left constituent of the rule was designed to look for 'As of' and 'On' before capturing the reporting date. Figure 2 shows the extraction task for a pattern of the form 'During 1-26 January 2003'; this expression is used to identify the period that the outbreak report is covering.

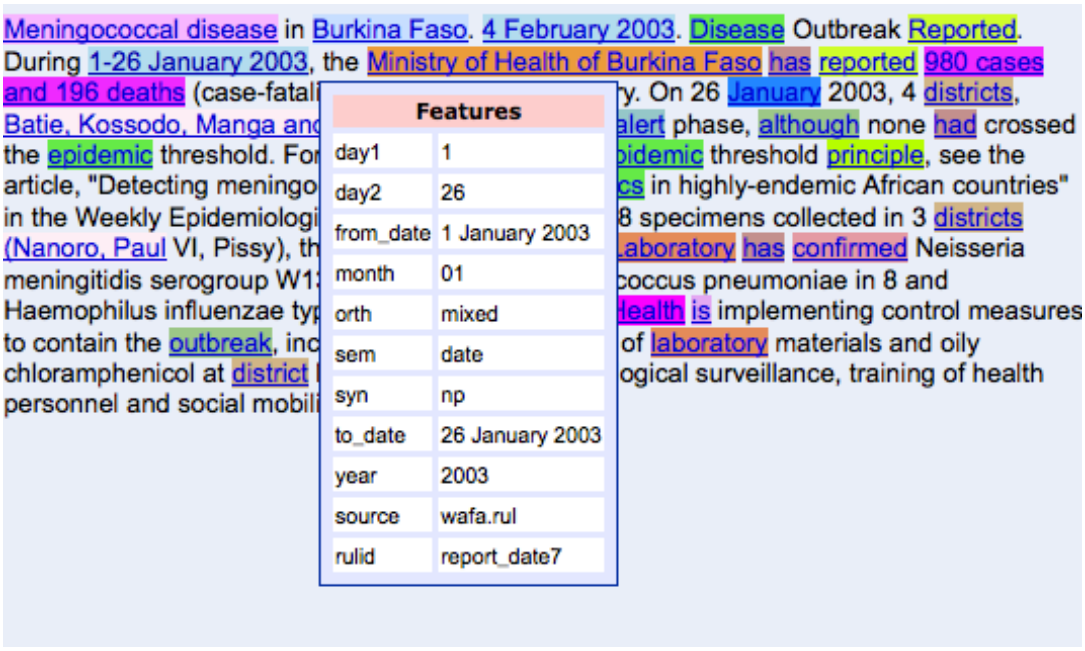


FIGURE 2: Announcement Date Extraction.

6.1.3. Country Name

There is a separate gazetteer for GeoName places, which occurs by default with the CAFETIERE system. When a country or city occurrence is found in the input text and matches an occurrence in the GeoName database, a phrasal annotation is made. Therefore, to extract the correct country of the outbreak, rules must be designed to classify the tagged locations.

A very common pattern is when the country name is preceded by the phrase 'Situation in'. The following rule has been created to capture this pattern:

```
[syn=NNP, sem=country_of_the_outbreak, type=entity, country=_c, rulid=country_name]
=>
[token="Situation"|"situation", sem!="collaboration"],[token="in"]
\
[syn =DT]?,
[sem>="geoname/COUNTRY", token=_c]
/;
```

However we found that a very common pattern involves indicating the country of the collaborating laboratory used to examine a virus. For example: 'Tests conducted at a WHO collaborating laboratory in the United Kingdom . . .'; therefore, phrases such as 'collaborating laboratory' and 'laboratory centre' were collected and added under the semantic class 'collaboration'. Now, whenever a country is mentioned after these words, it will not be extracted as the country of the outbreak.

6.1.4. Outbreak Name

In order to extract the name of an outbreak, it is essential to know as many disease names as possible. One way to do this is to incorporate a list of disease names, types and symptoms into our extraction system. The medical domain in particular is enriched with specific and generic dictionaries and terminology lists, including names and categories of diseases such as the ICD-10[6] disease lists. However, the emphasis of this project is to train the system to automatically identify disease names of various patterns and to be able to extract recently-discovered disease names and codes not found on the pre-fixed lists.

One of the main problems making the identification of disease names more complex than classic entities (such as person names and countries) is that they are in most cases written in lowercase. Therefore, there was a high need to recognize other features in the text. Another problem is when diseases are mentioned in the text but are not the outbreak itself, e.g. when the symptoms of an outbreak is a disease by itself. To avoid this situation, words indicating the occurrence of such a case are gathered and added in the gazetteer under the semantic class 'symptoms'.

Nouns that indicate disease types (usually mentioned after the disease name), such as virus, infection and syndrome, were all collected and added into the gazetteer under the semantic class 'disease_types'.

Many of the diseases are in the form of compound nouns where multiple words are used to describe one disease entity. A typical disease name may consist of the following:

sem(disease condition) sem(disease) sem(disease type)
'Disease conditions' is a new category created to cover all health conditions, such as 'acute', 'paralysis' and 'wild' that are used as disease descriptors. An example of this pattern is as follows:
'acute poliomyelitis outbreak'.

Another disease pattern is when the word 'virus' is attached to the end of the name, such as 'Coronavirus'.

Moreover, extraction rules have been designed to identify disease codes such as H1N1. The CAFETIERE system recognizes the word 'H1N1' as one token and assigns the value 'other' to the orthography feature because it contains characters and numbers; thus, the rules were designed based on this finding:

```
[syn=np, sem=Disease_code, key=__s, rulid=disease_code2] =>
\
[orth="other", token=__s],
[sem="disease_type", token=__s]
/;
```

The number of patterns for disease extraction we have designed and implemented total eighteen.

6.1.5. Affected Cities and Provinces

Not all of the cities, areas, provinces and states had been added into the GeoName database, resulting in them not being identified. This is either due to a transliteration problem or because they are not very well-known places. The problem was partially solved by studying the expression used to represent the locations in the text.

[6] ICD-10: The International Classification of Diseases standard.

In some texts, the names of the affected areas occur in expressions like '54 cases have been reported in the provinces of Velasco' and 'Cases also reported from Niari'. The first rule was designed to look for the following pattern:
sem(reporting_verbs) sem(preposition) sem(GeoName)

To identify any locations excluding country names, two categories from the GeoName database - geoname/PPL[7] and geoname/ADM2[8] - were used.

For locations not identified by the gazetteer, a number of rules have been designed to extract locations mentioned within explicit expressions. These expressions are usually used to express location and are followed by an indicating word after the location, such as 'city', 'province' and 'state':
sem(reporting_verbs) sem(preposition) Orth(capitalized) sem(areas)
This pattern conforms with expressions such as:
' . . . reported from the Oromiya region'.

Additionally, more rules were designed to capture groups of entities for situations where the outbreak hits more than one location. Identifying the name of the location seems to be the most challenging task, especially when the name of place is not identified using the gazetteer. Location phrases can take various forms and can occur anywhere in the text. The use of simple rules (finding location prepositions such as 'in' and 'from') may extract all the locations in the text, but this also may increase the number of false matches.

6.2. Relationship Extraction

The name of the reporting health authority

In the domain being studied in this project, the most interesting relationship we have found is the name of the health authority reporting the outbreak to the WHO. This relationship is a binary relationship of type "located in" (E.g. Ministry of Health, Afghanistan). This task is especially important because in some reports, the name of the country is not mentioned in the beginning but instead is implied in the name of the authority.

According to the texts under study, the reporting authority always takes the form of the relevant country's health authority, where the name of the health authority is adjacent to the country name. The most common form is:
sem(health authority) sem(preposition) sem(GeoName)

All the names that might refer to a health authority were collected and added to the gazetteer under the semantic class 'health_agency'. The most common authority reporting outbreaks was a country's 'ministry of health'; however, other names such as 'The Ministry of Health and Population' and 'The National Health and Family Planning Commission' were also found.

Pattern 1:
sem(health_agency) sem(preposition) sem(GeoName)
This will capture:
'Ministry of Health (MoH) of Egypt'

[7] PPL: "A city, town, village, or other agglomeration of buildings where people live and work", Source: The GeoNames geographical database: Available from:http://www.geonames.org/export/codes.html [Last Accessed: 12 June 2014]
[8] ADM2: "A subdivision of a first-order administrative division", Source: The GeoNames geographical database: Available from: http://www.geonames.org/export/codes.html [Last Accessed: 12 June 2014]

Pattern 2:
orth(DT) NNP(nationality) NP(health authority)
where DT refers to determiners such as 'the'. This pattern conforms with the following example:
'The Afghan Ministry of Public Health'.

Pattern 3:
sem(health authority) sem(punctuation) sem(health authority) sem(GeoName)
to capture:
'The National Health and Family Planning Commission, China'.

6.3. Event Extraction

Outbreak event

After examining 25 reports, it has been found that the patterns used to report an outbreak event are in the form of the number of victims of an outbreak reported by an authority. To avoid increasing the complexity of the events rules, the authority name is extracted in advance (relationship extraction).
Typically, the simplest event will be in the following form:

sem(GeoName) sem(reporting verbs) orth(CD) token ("cases")
This will capture a sentence in the following form:
'China reported 34 cases'.

However, as more texts are analyzed more constituents can be found, such as case classification and fatal cases. Therefore, before designing the events rules, key considerations have been taken into mind:

• Number of cases

The number of cases and deaths are usually in digit form, such as '134 cases'. Alternately, they can be in written form: 'five cases'. It has been also found that the form of 'twenty-five cases', where a dash is inserted between two numbers, is also used in some reports. Another issue related to extracting the numbers arises when the number consists of four or more units and a space is used after every three number units (e.g. '45 100'). To overcome the last problem, the number can simply be read as a whole string, '45 100', but if we want it to be saved as a proper integer value the following arithmetic calculation would solve the problem : total = "(+ (* _a 1000)_b))", CAFETIERE will interpret this calculation by multiplying the first number by 1000 then adding the second number.

For example, '45' is the first number token 'a' and '100' is the second token 'b', thus,
45 * 1000 = 45000
45000 + 100 = 45100.

• Case classification

Cases of infection from a disease are usually classified as either suspect, probable or confirmed cases to identify the degree of certainty of an outbreak. Those terms are known as 'case classification' and are often used in outbreak reports. Therefore, case classification has been added as a feature to the event extraction rules. All of the terms that fall under the case classification have been added to the gazetteer under the class 'case-classification'. Terms such as 'laboratory confirmed' and 'epidemiologically linked' are types of confirmed cases that have been added.

• Fatal cases

In addition to the typical classification of reported cases mentioned above, reports usually contain information about the number of fatal cases and deaths. To distinguish these cases from the others, their semantic class is 'fatal cases', and to distinguish the fatal but not dead from the deaths, the feature 'dead' will be used and will hold to values of either 'yes' or 'no' according to the terms used in the texts that describe the situation.

This simple event pattern 'China reported 34 cases' is very common; however, other similar patterns can be found:

'China reported 34 **new suspected** cases and 4 new deaths'.
'China reported 34 **new suspected SARS** cases and 4 new deaths'.

So to broaden the coverage of similar patterns, verbs that indicate the reporting such as 'reported', 'identified' and 'confirmed' were added along with their different tenses to the gazetteer. Their semantic class is 'reporting verbs'.

In addition to the active voice, passive patterns like
syn(CD) token ("cases") sem(haveverb +beverbs) sem(reporting_verbs)
are also used widely in outbreak reports; therefore, the verb groups such as 'have been' and 'has been' were added to the rules to capture the following type of pattern: '249 cases have been reported ...'.

Another example of an outbreak event pattern is:
orth(CD) sem(case_classification) token("cases") sem(preposition) syn(NN)

This will pick up sentences such as: '130 laboratory-confirmed cases of avian influenza'.
In addition to this, temporal and locative information may appear in different positions in the sentence or clause:
'Since 2005, 20 cases reported, 18 of which have been fatal'
'20 cases reported since 2005- 18 have been fatal'
'Of the 20 cases reported, 18 have been fatal since 2005'

More complex patterns can be found when both the temporal and locative information are mentioned in the same sentence:
"20 cases reported in Cambodia since 2005, 18 have been fatal."
Similarly, the phrase 'has reported' can occur anywhere in the reporting clause. The adjunct clause will be used to extract fatal cases such as number of deaths.

7. DISCUSSION

Even the texts chosen in this study belong to one domain, challenges caused by linguistic variation do exist. By linguistic variation we mean that different expressions may be used to deliver the same idea. Extracting information from texts can be achieved either by writing a few general patterns (which may lead to information being tagged under incorrect semantic classes) or by writing as many specific rules as possible (which will lead to an extensive workload by trying to write a rule for each pattern, even for those patterns that are rarely found in natural texts). Due to time constraints, both generic and specific rules were written to cover as many patterns for entities, relationships and events as possible.

Regarding the entities extraction, in the beginning we assumed that extracting the entities would be the most straightforward part of the project. This assumption has proven to be true for extracting the dates as they are always mentioned in the same way. This is also true for extracting the country of the outbreak. The only problem is with countries that are mentioned in the text but have no further reporting of disease outbreak. E.g.: 'Argentina and Peru have been notified of the cases that occurred earlier this month in Chile'.

The countries 'Argentina' and 'Peru' are not disease outbreak locations - 'Chile' is the outbreak location. So for this task, the work has been focused on the sentence level; countries mentioned in the first sentences are only captured if they conform to specific patterns, as it has been found that in the disease outbreak reports, the important information related to the actual outbreak event is always presented first and the secondary information is presented later.

Extracting the outbreak name was a relatively challenging task, as these do not conform to common patterns - even the orthography features are not obvious. Many diseases are named after the person who discovered them or after the location where they first appeared; this can cause confusion when extracting them. For example, the word 'Avian' was tagged by the gazetteer as the name of a location and not as the name of a disease. Some reports discuss the symptoms of an outbreak, which can be problematic if their sentence matches a pattern designed to capture an outbreak disease. All of these reasons have complicated the extraction process. Extracting the locations of an outbreak was very challenging task, especially for locations that were not tagged in the GeoName database; therefore, it was essential to discover as many expressions as possible.

Conversely, extracting the 'located in' relationship was relatively straightforward. This is because the reporting authorities have a limited number of patterns.

We initially assumed that events extraction would be difficult because the outbreak events are usually very long and consist of other information that may be extracted in advance; however, after closely examining and testing the patterns, it has been decided to treat each clause or sentence as a number of constituents indicating certain features. The longest event clause that can occur is when all the features are mentioned in the same sentence. By features, we mean that case classification, locative and temporal details and the number of cases are reported in the same sentence or clause. Other information that may appear in the event clause, such as the disease name and the country of the outbreak, is read only as a linguistic pattern that helps in constructing patterns but that is not extracted. This is because they are always extracted beforehand using separate rules. For example:
'130 laboratory-confirmed cases of human infection with avian influenza A(H7N9) virus including 31 deaths'. The information extracted is:

Number of cases = 130

Case classification = laboratory-confirmed

Number of fatal cases = 31

dead_cases = yes

The name of the disease will not be extracted and is only used to formalize one of the outbreak event patterns. In doing this, the extraction process will be facilitated as there is no reason to extract the same information again.

Most of the difficulties we encountered when designing the rules were due to rules order. The file containing the rules is ordered by featuring the rules for extracting entities at the beginning, followed by relationships and finally events. If a rule for entity extraction captures information from the clause containing the event information, the whole event will not be recognized. This problem was partially solved by defining the 'before' and 'after' constituents - the more conditions added, the more potential similarities between patterns will be avoided.

8. EVALUATION

The system was evaluated based on the scoring system used by the the Message understanding conferences [6]. The main findings of the MUCs are the measures of precision and recall, as well as the F-measure which is the average of precision and recall. Precision indicates how many of the elements extracted by the system are correct (accuracy), while recall indicates how many of the elements that should have been extracted were actually extracted (coverage).

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

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

In preparation for the evaluation, the sets of texts run through the system were also manually annotated. The evaluation process was based on a comparison of the manual extractions with the system's output. Elements extracted by the system were identified as:

- Correct (true positive): Elements extracted by the system align with the value and type of those extracted manually.
- Spurious (false positive or match): Elements extracted by the system do not match any of those extracted manually.
- Missing (false negative): The system did not extract elements that were extracted manually.
- Partial: The extracted elements are correct, but the system did not capture the entire range. For example, from the sentence "China today reported 39 new SARS cases and four new deaths", the system should extract the number of cases and deaths, but in this instance, it extracted only the number of cases. This case is a partial extraction and would be allocated a half weight, resulting in the coefficient 0.5. Another coefficient could be used to obtain more accurate results. For example, if the majority of an element is extracted, then a coefficient of 0.75 or higher can be used, but if only a small part of the element is extracted, a coefficient of 0.40 or less can be used. All the MUCs assigned partial scores for incomplete but correct elements [15].

Therefore, the measures of precision and recall can be calculated as follows:

$$\text{Precision} = \frac{\text{Correct + 0.5 Partial}}{\text{Correct + Spurious + Partial}} = \frac{\text{Correct + 0.5 Partial}}{\text{N}}$$

$$\text{Recall} = \frac{\text{Correct + 0.5 Partial}}{\text{Correct + Missing + Partial}} = \frac{\text{Correct +0.5 Partial}}{\text{M}}$$

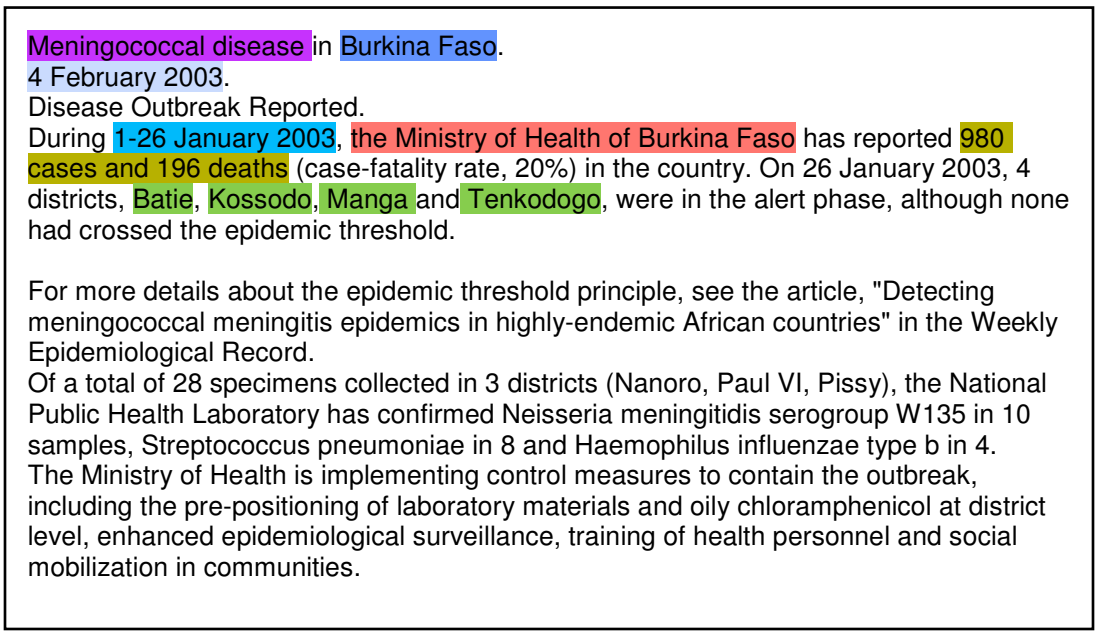
Where:
N = Total number of elements extracted by the system.
M = Total number of manually extracted elements.

8.1. System Evaluation Process

Ten texts new to the system were selected from the WHO website. In the training phase, 25 texts were chosen randomly. Summary reports of disease outbreaks in different countries were excluded from both the training and the testing sets because they typically are constructed differently and contain significantly different textual patterns.

The system tagged elements either because they were captured by the extraction rules or they matched a gazetteer entry. The main goal of this project was to test the extraction rules' ability to identify elements of the desired value and type. Elements tagged by the gazetteer consistently possessed the correct value and type and, if assigned a score, would receive the full score of 1. Therefore, it was decided to count only the elements captured by the extraction rules.

For example, a text was annotated manually to identify the elements that the system should extract (see Figure 3.


Meningococcal disease in Burkina Faso.
4 February 2003.
Disease Outbreak Reported.
During 1-26 January 2003, the Ministry of Health of Burkina Faso has reported 980 cases and 196 deaths (case-fatality rate, 20%) in the country. On 26 January 2003, 4 districts, Batie, Kossodo, Manga and Tenkodogo, were in the alert phase, although none had crossed the epidemic threshold.

For more details about the epidemic threshold principle, see the article, "Detecting meningococcal meningitis epidemics in highly-endemic African countries" in the Weekly Epidemiological Record.
Of a total of 28 specimens collected in 3 districts (Nanoro, Paul VI, Pissy), the National Public Health Laboratory has confirmed Neisseria meningitidis serogroup W135 in 10 samples, Streptococcus pneumoniae in 8 and Haemophilus influenzae type b in 4.
The Ministry of Health is implementing control measures to contain the outbreak, including the pre-positioning of laboratory materials and oily chloramphenicol at district level, enhanced epidemiological surveillance, training of health personnel and social mobilization in communities.

FIGURE 3: Manual Annotation.

Entities:

Outbreak name: Meningococcal disease
Country: Burkina Faso
Publish date: 4 February 2003
Report date start: 1 January 2003
Report date end: 26 January 2003
Outbreak locations: Batie, Kossodo, Manga and Tenkodogo

Relationship:

Reporting authority: Ministry of Health of Burkina Faso

Event:

Outbreak event: 980 cases and 196 deaths

The same text was run through the system (see Figure 4).

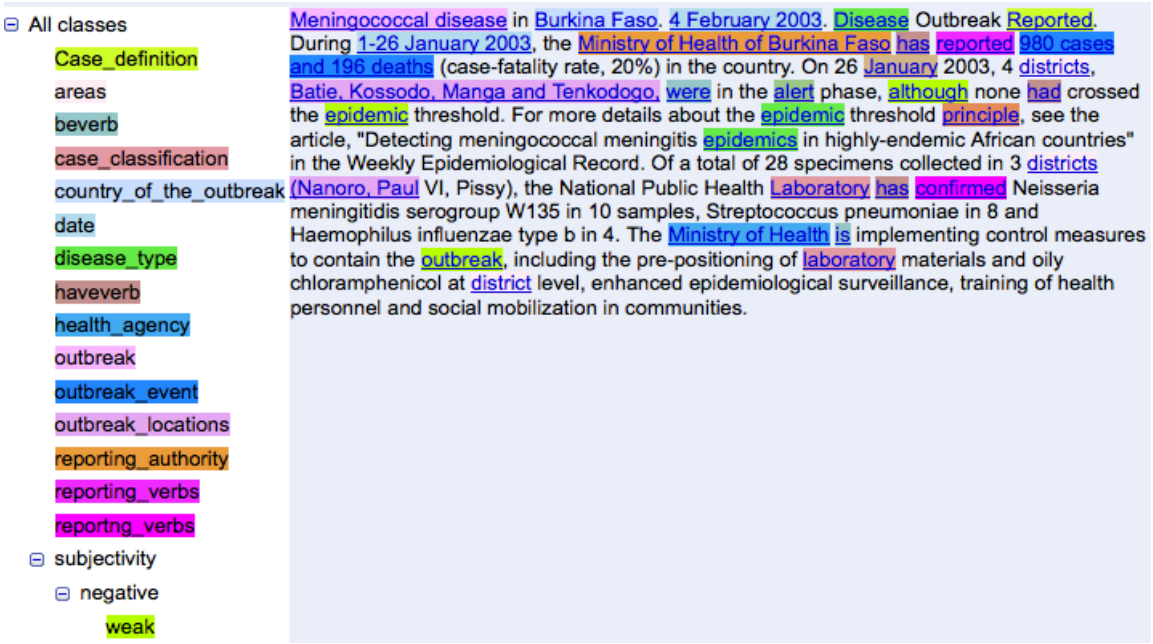


FIGURE 4: System Annotation.

As can be seen, the elements were correctly extracted and assigned to the appropriate type (class). The system extracted more elements than the manual process (gazetteer identification). Among the additional elements tagged by the system gazetteer in this particular example, the Ministry of Health is mentioned twice in the text. In the first instance, the phrase is followed by a country name. This pattern conforms to the reporting authority rules and therefore was tagged by the system. The second mention, however, did not follow a pattern recognized by any of the rules; therefore, it was tagged only by the gazetteer.

As well, additional tags which were not tagged by the gazetteer can be found and, in this case, are considered spurious elements. The same process of analysis was undertaken for both the train and the test sets.

9. RESULTS

	Entities			Relations			Events		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Average	0.95	0.85	0.90	0.80	0.80	0.80	0.90	0.92	0.91

TABLE 4: The evaluation metrics of the training corpus.

The system delivers a high level of performance when extracting outbreak events. Extremely high precision and recall were achieved in all events, high values were produced not only because the texts used were the actual training set, but also because the patterns utilized to extract the event were studied extensively in order to design additional rules for never-before-seen patterns. These additional patterns were predicted based on the knowledge that many tokens can separate adjacent pieces of information (the number of cases and deaths). The word ‘tokens’ here describes elements that can refer to temporal and location information, or to disease names.

The results of entities extraction also demonstrate extremely high performance. Recall is slightly lower than precision but is still considered high, with an average of 0.85.

The relationships were either correctly extracted or not extracted at all. Although the task of designing the relationships rules was relatively straightforward, it produced the lowest precision and recall, primarily because all the constituents of the relationship rule were made mandatory fields during the design phase, which prevented partial extraction.

	Entities			Relations			Events		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Average	1.00	0.75	0.88	0.70	0.70	0.70	0.90	0.80	0.85

TABLE 5: The evaluation metrics of the test corpus.

The most noticeable result is that all entities extracted from the test corpus were correct (Precision=1). The low recall indicated that extraction was performed with few errors, achieving high recall is generally more challenging than precision [8].

As in the training corpus, the relationship extraction had the lowest performance level. In addition to the possible cause discussed earlier, the rules did not take into account the new reporting patterns.

Event extraction for the test set also achieved a high level of performance even for new patterns. The results show an average precision of 90% and recall of 80%, reasonably high considering the complexity of the task.

It is necessary to determine the number of occurrences of each entity type in order to assess their level of difficulty. Not all the entities have the same frequency, necessitating accurate measurement of the performance of the rules.

Entity	Training set				Testing set			
	Correct	Partial	Spurious	Missing	Correct	Partial	Spurious	Missing
Published date	24				10			
Report date	14		1	6	5			1
Country	22			1	10			
Disease	23		2	1	6			4
Locations	28	2	3	17	15			11
Disease code	5				1			

TABLE 6: Number of occurrences of each entity type.

Table 6 shows that, in both the training and the testing sets, the location entity has the greatest number of missing elements. This result is not surprising because the extraction of locations was the most challenging task during the design phase. Locations can be mentioned anywhere in a report and do not conform to obvious patterns. In addition, unlike the other entities, locations can

be mentioned within a group of other locations as, for example, in the statement “Cases have also been reported in Larnaca, Famagusta, Nicosia and Paphos”. Problematically, in such patterns, the number of locations that can be mentioned within a clause may remain undetermined. In addition, the preceding and following sentences can present various patterns. In other epidemic surveillance specialised projects such as the HealthMap system, the location names considered very ambiguous entities achieving an F-score of 64%, the core idea behind health map is to identify disease outbreak locations using neural networks. This low performance was contributed to the problem of finding the definitive outbreak location among the geographic names mentioned in the text [16].Those factors make location one of the most difficult entities to handle within outbreak reports.

Another observation can be made about disease names. Although the design of the extraction rules for diseases was extremely challenging and required both a deep analysis of various disease names and the linguistic analysis of the context in order to prove that an entity actually was an outbreak, the results show highly accurate identification and few errors. These results are acceptable considering the difficulty of the task.

The overall system performance appears to provide better results comparing with those obtained from Proteus-BIO system which is also was evaluated using the MUC scoring system [12]. Table 7 shows the results of the test corpus in both the proposed system and Proteus-BIO system. As been discussed earlier, the proposed system is a rule-based system while the named entity recognition in Proteus-BIO is based on a machine learning algorithm called Nomen. The achievement of the proposed system is due to the use of hard-coded rules, where the patterns were studied extensively to discover powerful patterns and based on them, predict existed but not-yet-seen patterns.

Entity	The proposed system	Proteus-BIO system
Precision	86%	79%
Recall	75%	41%

TABLE 7: Comparative evaluation with Proteus-BIO system.

The extraction results, though, can be improved. In particular, the results for location and disease name entities could be bettered significantly by using up-to-date official datasets for location and disease names. Doing so would allow most effort to be focused on analyzing linguistic patterns rather than positing potential name structures and combinations.

10. CONCLUSION

The evaluation of the extraction rules yielded high precision and recall scores, close to those of state-of-the-art IE. The experiments were conducted independently with two subset corpora (the training and testing sets). The sets delivered similar system performance, although the training corpus had higher accuracy, particularly for relationship extraction. Event extraction, surprisingly, yielded to very high scores, the approach that helped in achieving such scores returns to the idea of looking what information digests that may form the event clause itself, so instead of only capturing the number of cases and deaths caused by the outbreak, other information was also included in the task such as case classification, fatality status, year and total numbers. Those constituents have helped in building many linguistic patterns that comprise the outbreak events.

It can be concluded that the rule-based approach has been proven capable of delivering reliable information extraction with extremely high accuracy and coverage results. This approach, though, requires an extensive, time-consuming, manual study of word classes and phrases.

11. FUTURE WORK

In the future, this research could be expanded in various directions. For instance, information about individual cases affected by an outbreak could be extracted, such as the gender, age, province, village and initial symptoms of a particular case. It would be useful to investigate how to use co-references in multiple sentences. In addition, the identification of location entities could be improved by combining the different levels of a location into a single relation; for example, Halifax, Nova Scotia, could be extracted as a location relationship. Finally, study should be directed toward reports on outbreaks affecting plants and animals.

12. REFERENCES

[1] J. Cowie, and W. Lehnert. (1996, Jan). "Information Extraction." Communications of the ACM. [On-line]. 39(1), pp. 80–91. Available: http://dl.acm.org/citation.cfm?id=234209 [Apr. 16, 2014].

[2] A. De Sitter, et al. "A formal framework for evaluation of information extraction." Technical report no. 2004-4. University of Antwerp Dept. of Mathematics and Computer Science, 2004. [On-line]. Available: http://wwwis.win.tue.nl/~tcalders/pubs/DESITTERTR04.pdf [Apr. 16, 2014].

[3] M. Moens. (2006). Information extraction: Algorithms and prospects in a retrieval context. [On-line]. 21. NewYork: Springer, 2006. Available: http://link.springer.com/book/10.1007%2F978-1-4020-4993-4 [Apr. 16, 2014].

[4] A. McCallum. (2005, Nov). "Information Extraction: Distilling Structured Data from Unstructured Text". ACM Queue. [On-Line]. 3(9), pp.48 -57. Available: http://dl.acm.org/citation.cfm?id=1105679 [Apr. 16, 2014].

[5] S. Acharya, and S. Parija. "The Process of Information extraction through natural language processing." International Journal of Logic and Computation. 1(1), pp. 40-51, Oct. 2010.

[6] R. Grishman, and B. Sundheim. "Message understanding conference - 6: A brief history." In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, 1996, pp. 466-471.

[7] H. Cunningham. "Information Extraction, Automatic." in Encyclopedia of language and linguistics, 2nd ed. vol. 5. Amsterdam: Elsevier Science, 2006, pp. 665-677.

[8] S. Sarawagi "Information extraction." Foundations and Trends Databases, 1(3), pp. 261-377, March. 2008.

[9] S. Esparcia, et al. "Integrating information extraction agents into a tourism recommender system," In Hybrid Artificial Intelligence Systems, vol. 6077. Springer Berlin Heidelberg, 2010, pp.193 – 200.

[10] J. Piskorski, and R. Yangarber. "Information extraction: Past, present and future." In Multi-source, multilingual information extraction and summarization, Part 1. Springer Berlin Heidelberg, 2013, pp. 23-49.

[11] Ahn, D. "The stages of event extraction" . In the Proceedings of the Workshop on Annotating and Reasoning about Time and Events, Sydney, Australia, 2006, pp.1-8.

[12] R. Grishman et al. "Information extraction for enhanced access to disease outbreak reports." BMC Bioinformatics, 35 (4), pp. 236–246, Aug. 2002.

[13] W.J. Black et al. “Parmenides Technical Report.” Internet: http://www.nactem.ac.uk/files/phatfile/cafetiere-report.pdf , Jan. 11, 2005 [Apr. 29, 2013].

[14] W.J. Black et al. “A data and analysis resource for an experiment in text mining collection of micro-blogs on a political topic.” In Proceedings of the Eighth International Conference on Language Resources and Evaluation, 2012, pp. 2083-2088.

[15] Maynard, D. et al. "Metrics for Evaluation of Ontology-based Information Extraction." In Proceedings of WWW 2006 Workshop on Evaluation of Ontologies for the Web”(EON), 2006.

[16] M. Keller et al. (2009, Dec.). “Automated vocabulary discovery for geo-parsing online epidemic intelligence.” Journal of Biomedical Informatics. [On-line]. 10(1): 385. Available: http://www.ncbi.nlm.nih.gov/pubmed/19930702, [Jun. 6,2014].

[17] W. Alshowaib. “Information Extraction.” Master thesis, University of Manchester, U.K., 2013.