Rule-based Information Extraction for Airplane Crashes Reports

Sarah H.Alkadi *alkadis@ksau-hs.edu.sa*
College of Science and Health Professions /Basic Science Department
King Saud bin Abdulaziz University for Health Sciences
Riyadh, 14611*, Saudi Arabia*

Abstract

Over the last two decades, the internet has gained a widespread use in various aspects of everyday living. The amount of generated data in both structured and unstructured forms has increased rapidly, posing a number of challenges. Unstructured data are hard to manage, assess, and analyse in view of decision making. Extracting information from these large volumes of data is time-consuming and requires complex analysis. Information extraction (IE) technology is part of a text-mining framework for extracting useful knowledge for further analysis.

Various competitions, conferences and research projects have accelerated the development phases of IE. This project presents in detail the main aspects of the information extraction field. It focused on specific domain: airplane crash reports. Set of reports were used from 1001 Crash website to perform the extraction tasks such as: crash site, crash date and time, departure, destination, etc. As such, the common structures and textual expressions are considered in designing the extraction rules.

The evaluation framework used to examine the system's performance is executed for both working and test texts. It shows that the system's performance in extracting entities and relations is more accurate than for events. Generally, the good results reflect the high quality and good design of the extraction rules. It can be concluded that the rule-based approach has proved its efficiency of delivering reliable results. However, this approach does require an intensive work and a cycle process of rules testing and modification.

Keywords: Information Extraction, Text Mining, NLP, Airplane Crashes, Rule-Based.

1. INTRODUCTION

With the advent of the internet, many documents are generated in a form that is not well-suited to automatic analysis by computers, making it difficult for humans to extract the information they need (Redfearn et al., 2006, McDonald et al., 2012). These obstacles make the field of information extraction one of the most attractive research areas for those seeking to help solve this kind of problem. IE can be effectively applied to various domains such as newswire services, biomedical reports, financial analysis, sport news, etc., with the ability to support different languages.

IE is regarded as the most important part of the pre-processing techniques involved in text mining (Kao & Poteet, 2006). IE phase is tasked to extract relevant data like entities, attributes, relations, and events. As such, IE systems rely on pattern-matching methods by analysing and finding general patterns, regular expressions, and the syntactic structures of a specific domain of information to be used in the extraction process. The extracted elements will then be stored in a database, ready for analysis by a data mining application.

Riloff and Lorenzen (1998) define IE systems as follows: "Information Extraction (IE) systems extract domain-specific information from natural language text. The domain and types of information to be extracted must be defined in advance. IE systems often focus on object identification, such as references to people, places, companies, and physical objects. Domain-specific extraction patterns (or something similar) are used to identify relevant information."

This definition contains some limitations in comparison to the present state of the field. An IE system is recognised to be an ideal system if it is “domain independent” or can be used on any domain rather than only on a specific one.

Additionally, Cowie and Lehnert (1996) give a still-appropriate definition, which is: "Information extraction isolates relevant text fragments, extracts relevant information from the fragments, and then pieces together the targeted information in a coherent framework. The goal of information extraction research is to build systems that find and link relevant information while ignoring extraneous and irrelevant information."

One limitation has been identified in the Cowie and Lehnert (1996) definition. Though the common focus in information extraction is on unstructured text processing, there are other unstructured sources that have not been covered, such as images and videos.

Because of the use of the CAFETIERE’s system and its limitation in accepting a limited number of formats, the project was restricted to a specific domain with only text-format input. The best definition that best explains the use of information extraction in this project is thus:
"Information Extraction is a technique used to extract relevant information such as entities, relations and events of a specific domain from large collections of documents, presenting it in a structured format for further analysis, where NLP techniques are applied and the criteria for extraction are pre-defined by the developer into a set of rules."

Therefore, this information extraction paper aims to develop deep understanding of information extraction through the development of an information extraction system for a specific domain. In doing so, an IE system is implemented to replace manual analysis procedures, and to satisfy the requirements for accurate results, and efficient and timely performance. A rules-based approach has been chosen for this project; this includes writing rules in the CAFETIERE IE system for airplane crash domains. Those rules derive information related to a user’s needs from a set of specific texts. These texts have been collected as inputs to be processed in several stages, after which the rules will be applied to extract certain information.

2. BASIC TYPES OF EXTRACTED ELEMENTS

There are four basic types of extracted elements that can be categorised as following (Feldman & Sanger, 2007; Redfearn et al., 2006):

1- **Entities:** Entities represent the basic building blocks of text structure that are easily determined in document collections such as people’s names, companies, geographic locations, products, dates, times, etc.
2- **Attributes:** Attributes identify the properties and features of the entities that have been extracted in the previous step, which might include a job title, a person’s age, etc.
3- **Facts and Relations:** These can be described as pre-defined relations between two or more entities as identified in the text. For example: “is employee of” (Steve Jobs, Apple): a relation between an employee and a company.
4- **Events:** These are considered the hardest elements to extract. Events represent the participation of entities in an activity or occurrence, and they are discerned by extracting several entities and the relationships among them. For example: launching a new product by a company announcement.

3. IE EVALUATION

For the purpose of improvement, the performance of Information Extraction systems needs to be evaluated quantitatively and qualitatively in order to ensure its efficiency by adopting good evaluation measures.

3.1. Preliminaries

Evaluating the IE system begins by scoring each template slot separately and then averaging the total results of slots to compute the overall score of an IE system (Jones and Galliers, 1996). A confusion matrix can be used to evaluate an IE system's performance, which is a common technique for counting system results. For extracted entities, four values need to be counted, which are:

1- **True positive (TP):** the number of relevant information items that have been extracted.
2- **False positive (FP):** the number of irrelevant information items that have been extracted.
3- **False negative (FN):** the number of relevant information items that should have been extracted.
4- **True negative (TN):** the number of irrelevant information items that have not been extracted.

All common measures such as recall, precision, and F-measure that have been introduced by MUCs can be easily calculated using the values of the confusion matrix (Sitter et al., 2004).

3.2. Classical Performance Measures

Over several MUCs, a consensus was reached among participants, such as research lab participants and sponsors, regarding how the evaluation process for information extraction systems would be measured. In MUCs, the extracted outputs were placed in hierarchical attribute-value structures called templates. Human annotator results for both training and test data were provided in a set of manual key templates. They were then compared against the automatic system outputs using an automatic scoring programme. Then, the scoring programme aligned the automatic extraction system templates with the manual key templates. The matching values were counted as correct; mismatching values were counted as incorrect; and template attributes with null values, which were not aligned with any key attribute, were counted as over-generation (Appelt & Israel, 1999; Appelt, 1999).

- **Precision (P) and Recall (R)**

Two common metrics were identified in the forum of MUC-3, which were precision (P) and recall (R). These are used, respectively, in measuring the accuracy and coverage of the system. This makes it possible to identify the values of both metrics for IE system outputs by providing the values of the total extracted entities manually [N key]; of the correct extracted entities [N correct]; and of the total possible responses of extracted entities [N response], as follows (Appelt, 1999; Sitter et al., 2004; Grishman, 1997):

$$P = \frac{N_{correct}}{N_{response}},$$

$$R = \frac{N_{correct}}{N_{key}}.$$

The recall calculates the ratio of the correct extracted information with the total number of information items that are extracted manually. Precision measures the ratio of correct extracted information with the total number of extracted information items presented in the text.

Undoubtedly, the optimization of both parameters is hard to achieve at the same time. In the case of optimizing a system for high precision, the extracted information will be highly relevant. In contrast, optimizing a system for high recall will cause the extraction system to see irrelevant information as relevant (Chinchor, 1992; Lewis, 1995; Appelt & Israel, 1999; Lehnert et al., 1994).

- **F-measure**

There is a need to combine recall and precision values in one metric, which was introduced in MUC-4 to enhance global comparisons among different systems. A statistic metric called F-measure was proposed in order to provide an overall score of the performance of the extraction systems. It is basically used to define the harmonic mean between both recall and precision. The recall (R) and precision (P) are weighted using the B parameter values to determine which one of them is more heavily weighted (Appelt & Israel, 1999; Appelt, 1999; Sitter et al., 2004).

For a given set of responses that are measured by recall (R) and precision (P) metrics, the F-measure is calculated as follows:

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}, \ 0 < \beta \leq 1.$$

Where: P = precision, R = recall

β = a factor that indicates the relative degree of importance assigned to recall and precision; when β equals 1, it means that equal importance is attributed to precision and recall. The metric is then called the (F1-measure) and is expressed as (Zhong et al., 2012):

$$F_1 = \frac{2 * precision * recall}{precision + recall}.$$

Thus, recall, precision, and F-measure are classified as the most frequently used metrics in evaluating information extraction systems. Those metrics are adopted in this paper.

4. DOMAIN SELECTION

Various areas can be nominated as domains for an IE project. For novices in this field, it is very challenging to pick a new domain and get optimal IE results. As mentioned previously, the domain of airplane crashes has been chosen for this study, as it includes many factors that support it as a domain of interest.

Airplanes represent one of the most important means of transportation today. They have made different parts of the world more accessible than ever before. Various air-travel incidents have thus gained a great deal of attention from newspapers and other media. Many newswires websites provide detailed reports about airplane crashes around the world.

The data extracted from airplane crashes can be used for various types of analysis. For example, the Aircraft Crashes Record Office (ACRO) and the European Aviation Safety Agency (EASA) are both tasked with presenting statistics on aviation accidents. They use news reports in their analysis, which contribute to the improvement of air transportation safety. Such extracted information can also be used to do the following:

1. Support air incident statistics to enhance aviation safety globally.
2. Analyse the main causes of airplane crashes.
3. Study the airlines and manufacturers with the best and worst records for air disasters.

These factors, along with the author’s personal interest in and knowledge of this domain, have gone into the decision to select it for this study.

Airplane incidents reports usually contain specific information such as crash date, crash site, airline, and so on, all of which represent the common pattern to be extracted.

4.1. Text Source Validation

News of major airplanes incidents, their damage and fatalities are recorded by various trusted news agencies, whereas minor airplane incidents that document the crashes of small personal airplanes or helicopters are not recorded as efficiently and elicit less interest. Consequently, only reports on major airplane crashes which include an airliner will be considered in this paper.

A wide range of specialised sites have been analysed, studied, and compared by developer to determine the most suitable text sources.

The main factors in choosing the nominated sites were:

- The minimal use of slang language.
- The sufficient quantity and quality of texts.
- The adherence of texts to the domain requirements.
- The use of common patterns to assist in the process of designing the extraction rules.

Thus, the nominated news agencies that were evaluated are:

Newswire	Website
Reuters	http://www.reuters.com
BBC	http://www.bbc.co.uk/
CNN	http://edition.cnn.com/
The Guardian	http://www.guardian.co.uk/
1001 Crash	http://www.1001crash.com

TABLE 1: The Nominated News Agencies for Airplane Crashes Reports.

The search engines of the nominated agencies were used to find news on the topic “airplane crash”. Such elements as writing style, formats, and details provided vary from one agency to another. This led to the decision to choose the site 1001 Crash as the source for text collection. The other news sources were discarded due to their limitations in meeting system requirements. Their texts either did not have the desired patterns, or included information about airplane crashes that was deemed too trivial.

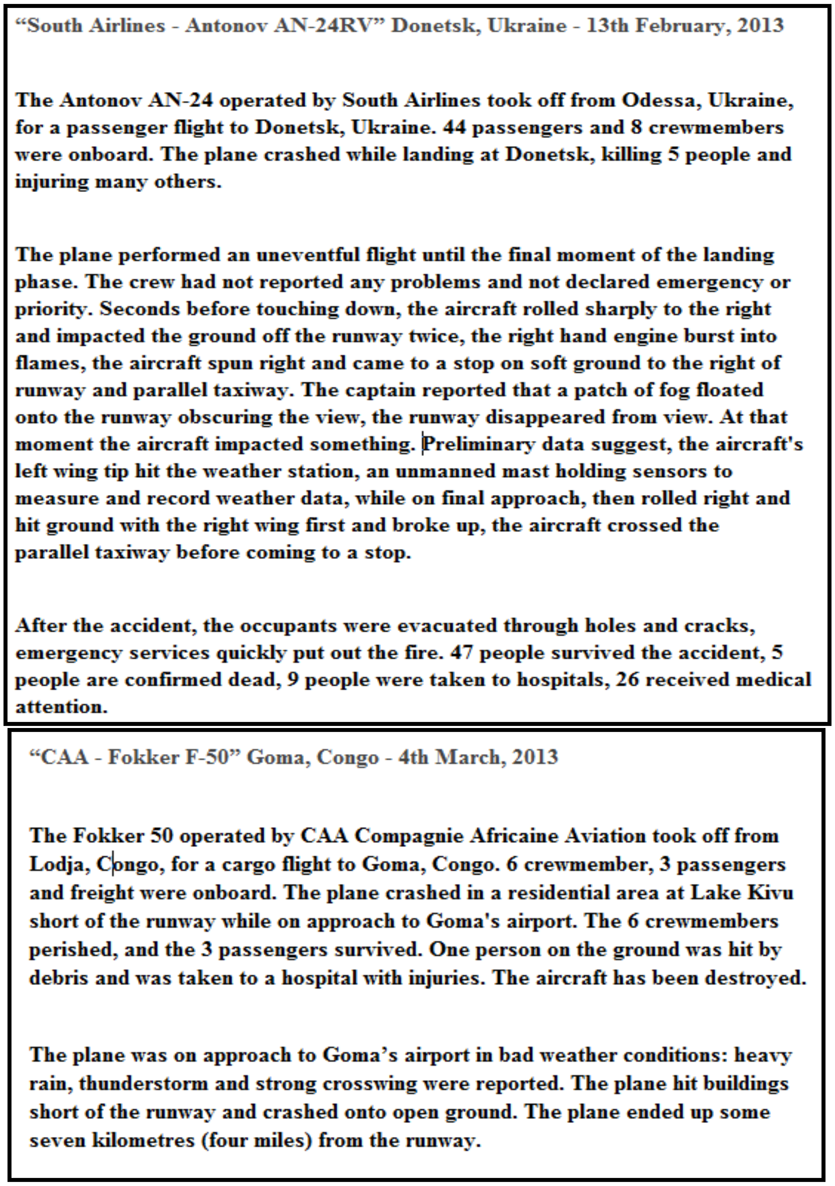
1001 Crash provides a wide range of worldwide airplane-crash news and statistics from 2000 to the present day with the following features:

- The site allows easy access to a huge number of texts.
- The texts frequently include information that satisfies the domain requirements, such as crash site, departure, destination, airline, etc.

It is worth noting that the 1001 Crash reports had to be converted into plain text before they were uploaded to the system.

4.2. Text Analysis

Text samples from 1001 Crash were randomly selected for analysis. The chosen texts reported on incidents that occurred between 2011 and 2015. A set of sample texts, two of which are listed below, was used as a working corpus for designing the extraction rules.



"South Airlines - Antonov AN-24RV" Donetsk, Ukraine - 13th February, 2013

The Antonov AN-24 operated by South Airlines took off from Odessa, Ukraine, for a passenger flight to Donetsk, Ukraine. 44 passengers and 8 crewmembers were onboard. The plane crashed while landing at Donetsk, killing 5 people and injuring many others.

The plane performed an uneventful flight until the final moment of the landing phase. The crew had not reported any problems and not declared emergency or priority. Seconds before touching down, the aircraft rolled sharply to the right and impacted the ground off the runway twice, the right hand engine burst into flames, the aircraft spun right and came to a stop on soft ground to the right of runway and parallel taxiway. The captain reported that a patch of fog floated onto the runway obscuring the view, the runway disappeared from view. At that moment the aircraft impacted something. Preliminary data suggest, the aircraft's left wing tip hit the weather station, an unmanned mast holding sensors to measure and record weather data, while on final approach, then rolled right and hit ground with the right wing first and broke up, the aircraft crossed the parallel taxiway before coming to a stop.

After the accident, the occupants were evacuated through holes and cracks, emergency services quickly put out the fire. 47 people survived the accident, 5 people are confirmed dead, 9 people were taken to hospitals, 26 received medical attention.

"CAA - Fokker F-50" Goma, Congo - 4th March, 2013

The Fokker 50 operated by CAA Compagnie Africaine Aviation took off from Lodja, Congo, for a cargo flight to Goma, Congo. 6 crewmember, 3 passengers and freight were onboard. The plane crashed in a residential area at Lake Kivu short of the runway while on approach to Goma's airport. The 6 crewmembers perished, and the 3 passengers survived. One person on the ground was hit by debris and was taken to a hospital with injuries. The aircraft has been destroyed.

The plane was on approach to Goma's airport in bad weather conditions: heavy rain, thunderstorm and strong crosswing were reported. The plane hit buildings short of the runway and crashed onto open ground. The plane ended up some seven kilometres (four miles) from the runway.

FIGURE 1: Text samples from 1001 Crash Website.

Looking at the texts above, obvious texts' characteristics can be realized:

- **The length of the text:** It differs due to differences in the reporting of these two air incidents. In this project, both short and long texts were considered, with the restriction of one hundred words as a minimum text length. Short texts, as can see above, indicate things like the crash site, crash date, departure, destination, and a summary of casualties and damage. Long texts include the basic information and more details about casualties, damage, and stages of a crash, along with some comments from affected individuals.

- **The title of the text:** It includes summaries of airplane type, airline, destination, and crash date. The airplane type in the title appears in a different form from that used in the text body, even though both refer to the same airplane type, such as "Boeing 747-400BCF" and "Boeing 747-400". The title presents the more specific rendering, which does not change any fact in the context of crash reports. Consequently, the first three elements of the text title were skipped due to their repetition in the body of the text in a more suitable structure. In terms of the crash date, it is only mentioned once, in the title. Therefore, only the crash date was extracted from the title.

- **The indicators of extracted elements:** Extracting entities was conduct first and was fairly easy, due to their clear patterns, such as numbers or keywords. For example, the title abbreviations “Mr.” or “Mrs.” appearing before a capitalised word help in recognising a person’s name. Moreover, words like “airlines” or “airways” indicate the airliners. However, some entities are more challenging to extract such as crash site. Relations and events, for their part, are built from a set of extracted entities. This poses challenges in writing rules for events and relations because of their correlation to the task of extracting entities. Additionally, the difficulty of extraction is higher with events than it is with relations.

- **The format of chosen texts:** There are some orthography issues that need to be considered when designing the rules — for example, the format errors in a phrase like “Atlasjet airline”, in which the word “airline” is not written correctly in capitalized form like "Atlasjet Airline". Another example is the (past-tense) verb “took off”, which has been written incorrectly in some texts as “took of”, or typographical errors in prepositions, like “en” instead of “on”. Therefore, it is important to be conscious of these issues and not to trust grammar or spelling fully as they appear in a text, as the text might include errors.

- **The writing style:** The texts follow the American English style, which is noticeable in some spellings — for example, words such as “airplane”, “kilometer”, or ‘meter”. This style needs to be taken into account throughout the design stage.

Therefore, the order of entities rules depends on the complexity level of these rules. The rules for extracting information related to crash sites must thus be written last in the extraction-rules chain of entities. Relations are the next in this chain which is used to identify a single type of information such as the relation between aircraft and its airliner. In terms of events, most events have more information with high levels of complexity. As, it is quite difficult to priorities the order of event rules early in the early development process.

Issues related to orthography might be handled in one of two ways. The developer could simply accept the previous errors during the rule design; or, alternatively, the developer could use supported applications such as spellchecker or grammar checkers during the pre-processing stages to resolve these issues early in the process. In the case of American English writing style, rules will be designed to accommodate them.

4.3. Entities, Relationships and Events Identification

After a deep study of the selected texts, the main information to be extracted was identified. The information included a group of entities, relations, and events that can be categorised as follows:

Element Type	Example
Entities	Crash site Crash date and time Departure and destination Aircraft, airline and manufacturer

	Flight purpose Passenger numbers Crash's geographical dimension.
Relations	The relation between the aircraft and its operating airline.
Events	Crash announcement. Crash casualties, which include those injured, killed, and surviving. Crash damage.

TABLE 2: The Examples of Each Type of Extracted Elements.

5. INFORMATION EXTRACTION SYSTEM

CAFETIERE stands for Conceptual Annotations for Facts, Events, Terms, Individual Entities, and RElations. It is a web-based IE system developed by the National Centre for Text Mining at the University of Manchester for text mining and analysis tasks. It uses the knowledge engineering approach to perform the information extraction function. It is used in this project. It follows the Apache UIMA (Unstructured Information Management Applications) framework[1].

5.1. Information Extraction Stages

The input documents are processed in several pre-processing stages before extraction rules are applied (Black et al., 2005). These stages are shown in the following figure:

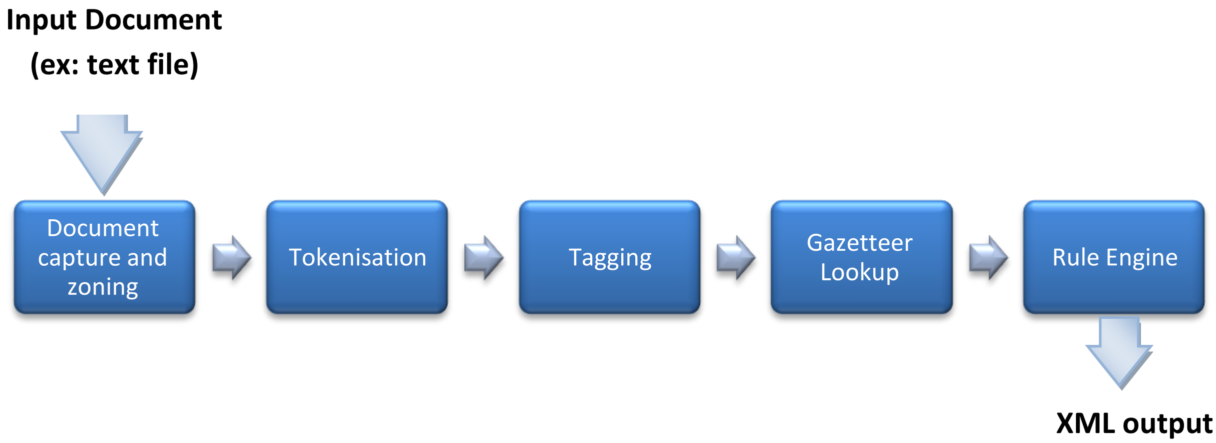


FIGURE 2: Information extraction stages of the CAFETIERE system.

- **Document Capture and Zoning**
  This includes capturing and zoning documents by converting the document's native markup to the structural markup of CAFETIERE's annotation scheme. The input text can be in plain text, HTML, or SGML formats, which will be processed by first separating the front matter of text from its body, then splitting the text into paragraphs.

- **Tokenisation**
  At this stage, the text of each paragraph will be partitioned into basic units such as numbers, words, punctuation, and symbols. Those units represent the tokens that comprise the whole text. When the token is identified, a token object is created, involving both its string representation and specific attributes, such as the token's position in the

[1] Apache UIMA project available at: http://uima.apache.org

text, as well as its orthography code. The extraction rules should thus be designed to be consistent with the tokeniser workflow in processing the text fragments. For example, the time format 5:20 GMT is treated by the CAFETIERE system as four tokens consisting of two digits, punctuation, and the time-zone acronym, all of which are extracted token by token.

- **Part-Of-Speech Tagging (POS):**
  The POS tags the extracted tokens into categories based on the semantic content of the words. Nouns, verbs, adverbs, common nouns, adjectives, and prepositions are the most common tags. The Penn-Treebank II tags[2] have been used in the POS tags in CAFETIERE.

- **Gazetteer Lookup:**
  At this stage, words will be labeled according to their semantic class if they are stored in the specific dictionaries — gazetteers. The gazetteer lookup is a database that includes the semantic categories of relevant words and phrases for specific domains.

- **Rule Engine:**
  In the final stage, the rule engine extracts named entities, relations, and events. After the rules are applied, the extracted elements will be stored with their semantic and syntactic features in the database for further analysis. Those elements can be viewed later in an annotation browser or exported as an XML file.

The developer will be responsible for implementing the final stage. This stage includes designing a set of extraction rules to be applied to pre-selected text collections. The rules will be modified and tested again in a cycle until the results match the requirements. The first four stages have already been implemented in the CAFETIERE system (Black et al., 2005).

5.2. Rule Notation

The general form for designing rules in the CAFETIERE system is the following:

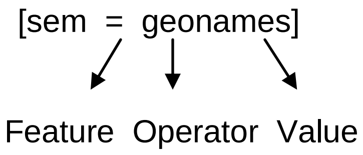
```
Phrase => context B \ Constituents/context D;
```

FIGURE 3: The Form of Extraction Rule.

Where (Phrase) represents the phrase or word to be extracted and (Constituents) include the text elements of the phrase.

Contexts B and D are the neighbouring text elements of constituents and should appear immediately before and after, respectively. There might be one or more left and right contexts, or these might not exist at all. Rules without B and D parts are called context-free, while rules with either right or left contexts are called context-sensitive. The symbols "\", "/", and at least one constituent are compulsory (Black et al., 2005; Black, 2007; Black 2013).

The rule elements "phrase", "constituents", and "context" are expressed as a set of "feature:operator:value" forms which are expressed as in the example the below:



[2] Penn-Treebank tags available at: http://www.cis.upenn.edu/~treebank/

They are enclosed within square brackets and separated by commas. Those elements are clarified in short as follows:

1. **Feature:** involves the attribute names appearing in the text units. Syn, sem, orth, and id are the most frequently used features. Syn is used to identify the part of speech, such as CD, IN, DT, NNP, NN, etc. The sem attribute is used to label the semantic classes of words or phrases. The semantic classes might already exist in the gazetteer or have been defined by other rules.

2. **Orth:** indicates the orthography of a token, using a set of codes such as uppercase, lowercase, etc. A unique identifier for the token in the text is determined using the id attribute.

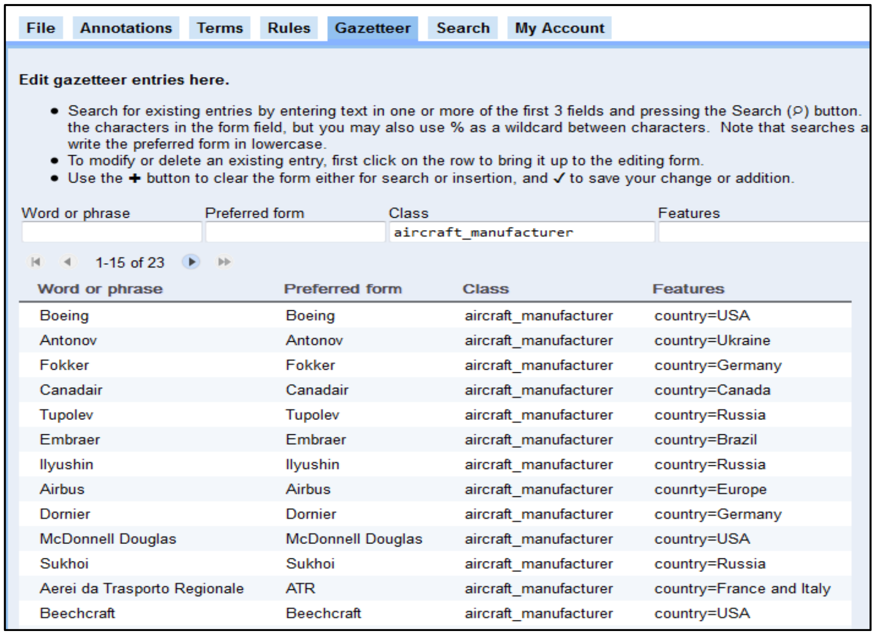
3. **Operator:** can be one of the following symbols: {=, !=, >, <, >=, <=, ~}; these are applied to the text elements' attributes. Each of which denotes a well-known function with a common meaning. In case of "~", it represents a pattern match operator.

4. **Value:** can be associated with one of the following three options: a literal (quoted or unquoted string or number), a pattern, or a variable. The pattern of regular expression is used to expand the scope of the rule which is considered to be an alternative to the literal value. The variables can also be used.

5.3. Gazetteer

An external resource called Gazetteer is provided by the CAFETIERE system for tagging words in domain texts, in addition to separating them into categories depending on their semantic content. For each domain, a special gazetteer is created by the developer.

CAFETIERE presents default gazetteers such as geoname, which includes cities and countries names, and many more. The developer might thus either settle for the default dictionaries, or create his/her own new, domain-specific gazetteer. The following figure shows an example of airplane crash gazetteer entries which have been created by the developer:



Edit gazetteer entries here.

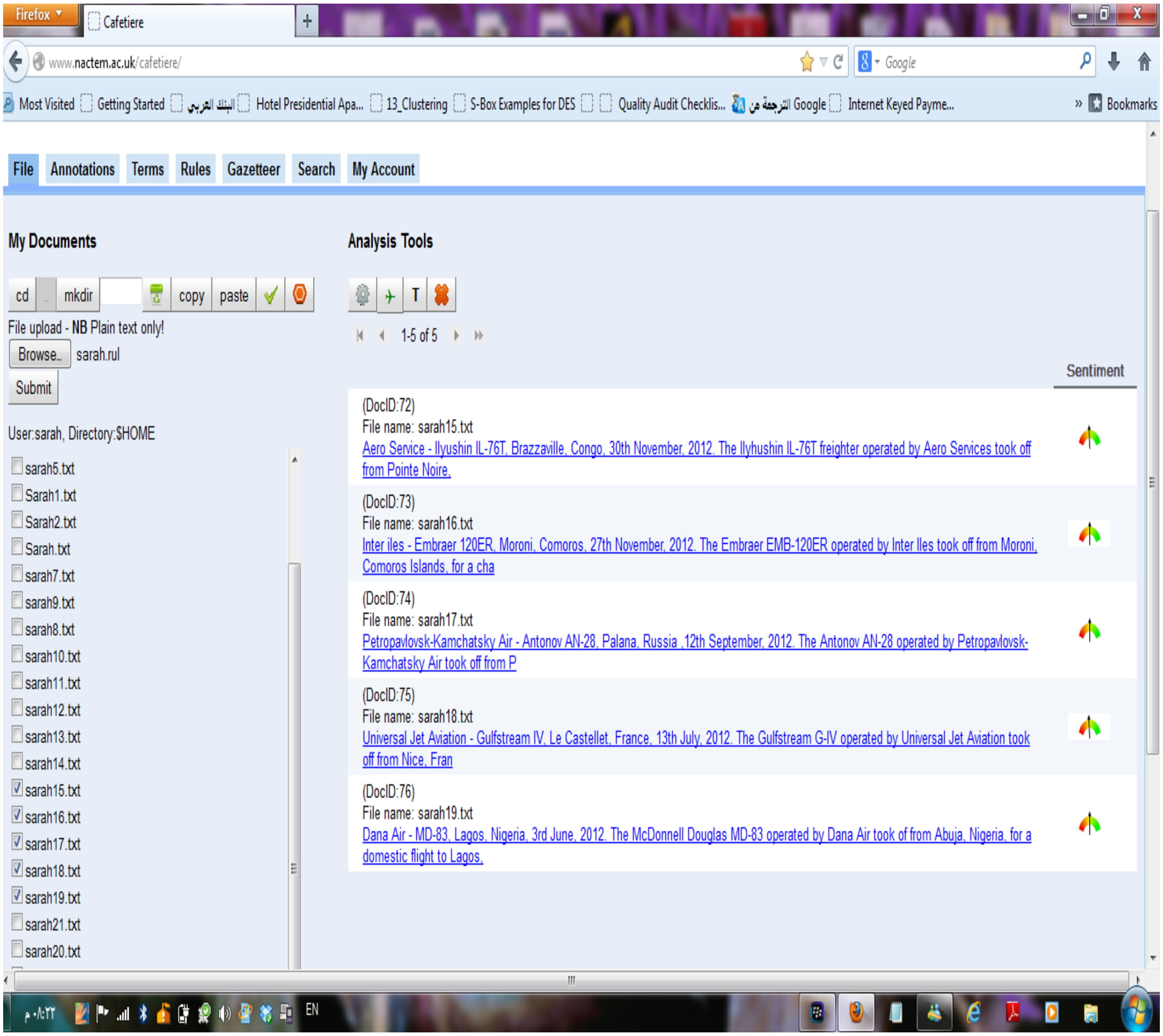
- Search for existing entries by entering text in one or more of the first 3 fields and pressing the Search (⚲) button. the characters in the form field, but you may also use % as a wildcard between characters. Note that searches a write the preferred form in lowercase.
- To modify or delete an existing entry, first click on the row to bring it up to the editing form.
- Use the ✚ button to clear the form either for search or insertion, and ✓ to save your change or addition.

Word or phrase:
Preferred form:
Class: aircraft_manufacturer
Features:

Word or phrase	Preferred form	Class	Features
Boeing	Boeing	aircraft_manufacturer	country=USA
Antonov	Antonov	aircraft_manufacturer	country=Ukraine
Fokker	Fokker	aircraft_manufacturer	country=Germany
Canadair	Canadair	aircraft_manufacturer	country=Canada
Tupolev	Tupolev	aircraft_manufacturer	country=Russia
Embraer	Embraer	aircraft_manufacturer	country=Brazil
Ilyushin	Ilyushin	aircraft_manufacturer	country=Russia
Airbus	Airbus	aircraft_manufacturer	countrty=Europe
Dornier	Dornier	aircraft_manufacturer	country=Germany
McDonnell Douglas	McDonnell Douglas	aircraft_manufacturer	country=USA
Sukhoi	Sukhoi	aircraft_manufacturer	country=Russia
Aerei da Trasporto Regionale	ATR	aircraft_manufacturer	country=France and Italy
Beechcraft	Beechcraft	aircraft_manufacturer	country=USA

FIGURE 3: Gazetteer's Entry Examples.

For example, trigger words like “airline” is entered in the gazetteer as follows:
airline: instance=airline, class=airline_indicator



My Documents

File upload - NB Plain text only!

sarah.rul

User:sarah, Directory:$HOME

□ sarah6.txt
□ Sarah1.txt
□ Sarah2.txt
□ Sarah.txt
□ sarah7.txt
□ sarah9.txt
□ sarah8.txt
□ sarah10.txt
□ sarah11.txt
□ sarah12.txt
□ sarah13.txt
□ sarah14.txt
☑ sarah15.txt
☑ sarah16.txt
☑ sarah17.txt
☑ sarah18.txt
☑ sarah19.txt
□ sarah21.txt
□ sarah20.txt

Analysis Tools

1-5 of 5

Sentiment

(DocID:72)
File name: sarah15.txt
Aero Service - Ilyushin IL-76T, Brazzaville, Congo, 30th November, 2012. The Ilyhushin IL-76T freighter operated by Aero Services took off from Pointe Noire,

(DocID:73)
File name: sarah16.txt
Inter iles - Embraer 120ER, Moroni, Comoros, 27th November, 2012. The Embraer EMB-120ER operated by Inter Iles took off from Moroni, Comoros Islands, for a cha

(DocID:74)
File name: sarah17.txt
Petropavlovsk-Kamchatsky Air - Antonov AN-28, Palana, Russia ,12th September, 2012. The Antonov AN-28 operated by Petropavlovsk-Kamchatsky Air took off from P

(DocID:75)
File name: sarah18.txt
Universal Jet Aviation - Gulfstream IV, Le Castellet, France, 13th July, 2012. The Gulfstream G-IV operated by Universal Jet Aviation took off from Nice, Fran

(DocID:76)
File name: sarah19.txt
Dana Air - MD-83, Lagos, Nigeria, 3rd June, 2012. The McDonnell Douglas MD-83 operated by Dana Air took of from Abuja, Nigeria, for a domestic flight to Lagos,

6. PROJECT METHODOLOGY

According to project workflow, the rules are designed first, then implemented in domain-specific texts. Those rules are then tested in an iterative process to ensure compatibility with user requirements. Consequently, a combination of prototype and waterfall methodologies will be applied to be the current project method, as is shown in the following figure:

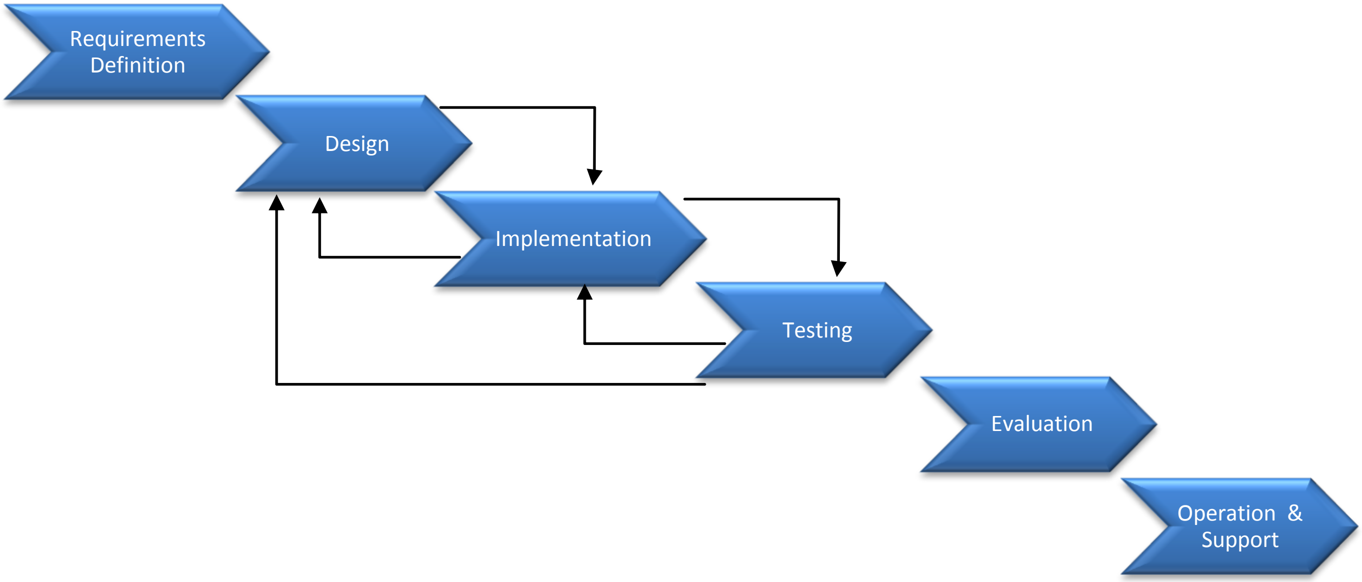


FIGURE 5: The adaptive development methodology for the IE system

All of the project phases will be defined using the waterfall approach, while the prototype approach will be applied to the design, implementation and testing phases. Consequently, the basic requirements, which are entities, relations, and events, will be set in advance, reflecting the workflow of the first approach. The written rules will be modified and tested iteratively until their correctness is ensured and outputs are satisfied.

7. DESIGN, IMPLEMENTATION AND TESTING

The development process of the airplane crash domain rule includes concerning about main features of the chosen texts. Important features of domain texts were counted while designing the extraction rules. They represent the linguistic patterns that appeared in the main texts that assisted the extraction:

1. Attribute: The text element is characterised by a set of attributes such as its
2. morphological root, its part of speech, its semantic class, linguistic attributes such as tense or determiner, or by its orthography, all of which are used to identify the text elements precisely. For example, the aircraft name “Airbus” might be defined using it orthographic feature, “Capitalised”.
3. Constituent Elements: The text element might be defined by its prefix or suffix constituent, which are known as the “constituent elements”. For example, “60 minutes” is identified as time when the number is followed by the time measurement unit “minutes”.
4. Context: The elements of text can be recognised according to the context in which they appear. The context is usually provided by the elements either directly before or after the word to be extracted. For example, the identification of an airline company such as “Jet One Express” can be determined by prior words such as “operated by”, which are followed directly by the airline company’s name, with its initial letter capitalised.
5. Pre-extracted elements: A text element can be recognised according to pre-extracted elements. This case reflects the importance of rule order. For example, to extract the relation between an aircraft and its airline, both the airline and aircraft entities must have been recognised earlier.
6. Co-referents: A text element can be defined by its co-referents. For example, when an airplane company has been identified successfully, it is possible for it to appear again in the text in different form, such as its referred pronoun. However, the CAFETIERE system,

which is used in this project, does not support coreference extraction. Therefore, text elements are not identified by their coreferents here.

Moreover, there were additional entries to the system gazetteer by collecting domain-specific vocabularies to be entered then and categorized under an appropriate semantic class.

7.1. Rules for Entity Extraction

As mentioned, the information extraction procedure starts with the extraction of entities. Entities are thus of great importance in the extraction chain, and their rules must be written according to their priority. The rules governing entities will be discussed in detail in the following section.

7.1.1 Rules for Extraction Crash Date and Time

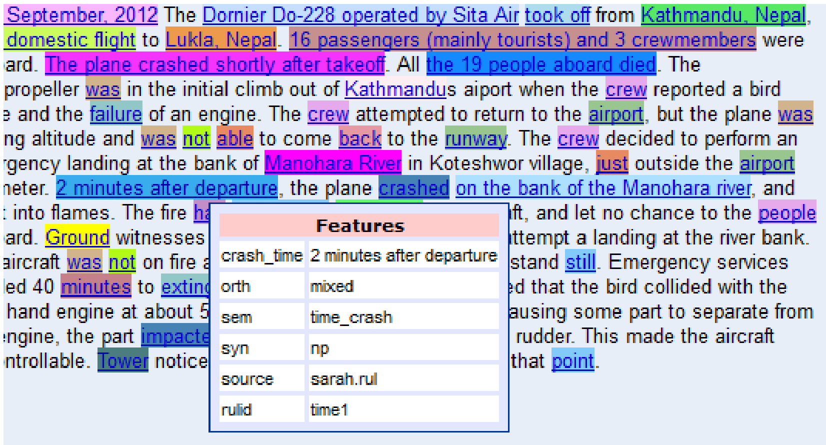
Reporting of crash date and time is counted as an important element when choosing the text source. In 1001 Crash texts, the date of any airplane crash is indicated clearly in the title following one main format, as follows: “DDth M, YYYY”. Here, day and year are represented using the numeric format, while the month is expressed in the textual format, as in “5th November, 2003”.

# Simple date rule

```
[syn=np, sem=date, type=entity,Crash_Date=__t, Day=_day, Month=_mnth, Monthno=_mno,
Year=_year, rulid=date1] =>
\
[syn=CD, orth=number, token=__t, token=_day],
[token="th"|"st"|"nd"|"rd" , token=__t],
[sem="temporal/interval/month", monthno=_mno, key=_mnth ,key=__t],
[token=","]?,
[syn=CD,orth=number, token~"19??"|"20??", token=__t, token=_year]/;
```

According to the previous rule, month names are stored by default in the gazetteer database under the class "temporal/interval/month". The day might be extracted as single- or double-digit numbers, such as 3 or 23, followed by the letters “th”. This pattern refers to ordinal numbers, which include numbers with the suffixes -th, -st, -nd or -rd. The year is identified by four-digit numbers, such as 2005.

The time of a crash is expressed in one simple format, with little variation. Time is indicated using the measurement units minutes, hours, or seconds, calculated in relation to the airplane’s departure, as in “2 minutes after departure”. The previous time measurement units are considered to be trigger words in extracting the time, as shown in the figure below:



September, 2012 The Dornier Do-228 operated by Sita Air took off from Kathmandu, Nepal,
domestic flight to Lukla, Nepal. 16 passengers (mainly tourists) and 3 crewmembers were
ard. The plane crashed shortly after takeoff. All the 19 people aboard died. The
propeller was in the initial climb out of Kathmandus aiport when the crew reported a bird
e and the failure of an engine. The crew attempted to return to the airport, but the plane was
ng altitude and was not able to come back to the runway. The crew decided to perform an
rgency landing at the bank of Manohara River in Koteshwor village, just outside the airport
neter. 2 minutes after departure, the plane crashed on the bank of the Manohara river, and
into flames. The fire ha ft, and let no chance to the people
ard. Ground witnesses ttempt a landing at the river bank.
aircraft was not on fire a stand still. Emergency services
ed 40 minutes to exting ed that the bird collided with the
hand engine at about 5 ausing some part to separate from
engine, the part impacte rudder. This made the aircraft
ntrollable. Tower notice that point.

Features	
crash_time	2 minutes after departure
orth	mixed
sem	time_crash
syn	np
source	sarah.rul
rulid	time1

FIGURE 6: Crash time extraction**.**

7.1.2 Rule for Extraction of Airplane Type

Various airplanes are made by different manufacturers. In the chosen texts source, the airplane type always starts with the name of the manufacturer, represented as a capitalised word followed by a set of letters or numbers and punctuation marks, as in" Tupolev Tu-154B-2". The well-known manufacturers in the aviation industry, along with their home countries feature, have been added to the gazetteer to assist during the extraction process. For example, the Boeing company is entered into the gazetteer under the class (airplane_manufacturer) and with the feature (country=USA).

Additionally, some phrases indicate the airplane’s purpose within the type, i.e. whether it is for passengers, cargo, etc. Apart from that, the number of textual units within each phrase varies, such as the Sukhoi Superjet 100 (SSJ100), the Boeing 737-210C/Adv (combi aircraft). This led to the design of multiple rules to handle the various patterns.

All the developed rules include the regular post context to determine the end of the phrase precisely.

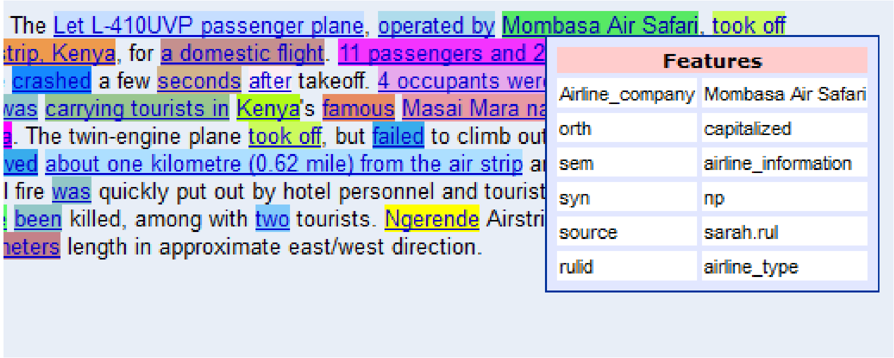
7.1.3 Rules for Airline Extraction

Different formats are used to provide the airline information such as: South Airlines, Sukhoi, Mombasa Air Safari, etc. Many signs ease the process of extraction which are:

1. The airline is always presented with initial uppercase letters, which can be recognised using the corresponding orthographic feature.
2. The airline indicator words such as Airlines, Airways, etc., which were entered in the gazetteer and considered while designing the rules. In some cases, the airline indicators are not mentioned within the airline phrase, which makes the extraction of airline difficult. The use of semantic class “operateverb” as a pre-context was a perfect solution to prevent any potential failure, and assists in the extraction.

Additionally, more restrictions can be added to avoid incorrect extractions by using the common pattern of post-context taken from the texts. The domain gazetteer’s new class of “departverb”, besides the built-in semantic classes like “beverb”, which covers the regular verb tense of phrase, and some punctuation, represent the regular post-context pattern.

The following screenshot shows the airline extraction:



The Let L-410UVP passenger plane, operated by Mombasa Air Safari, took off
trip, Kenya, for a domestic flight. 11 passengers and 2
crashed a few seconds after takeoff. 4 occupants were
was carrying tourists in Kenya's famous Masai Mara na
a. The twin-engine plane took off, but failed to climb out
ved about one kilometre (0.62 mile) from the air strip ar
l fire was quickly put out by hotel personnel and tourist
been killed, among with two tourists. Ngerende Airstri
neters length in approximate east/west direction.

Features	
Airline_company	Mombasa Air Safari
orth	capitalized
sem	airline_information
syn	np
source	sarah.rul
rulid	airline_type

FIGURE 7: Extraction of airline information**.**

7.1.4 Rules for Extraction of Flight Purpose

Flights are operated for various purposes, such as for to carry passengers, deliver cargo, etc. Three different formats were defined from the texts source as follows:

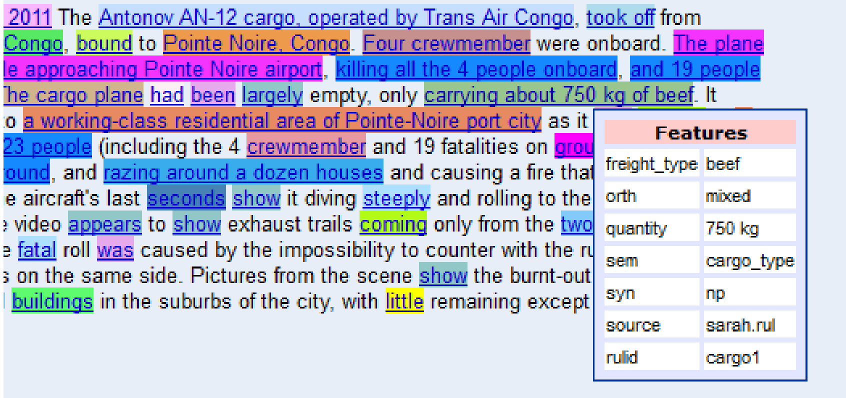
"for a passenger flight"
"was performing a cargo flight"

"The Cargo Plane"

Those formats led to the design of three types of extraction rules. The first type captures the majority of flight-purpose mentions in the texts. It uses the orthographic features and trigger words such as "flight" to fetch the flight purpose accurately. The other types of extraction rule uses the gazetteer classes, the common pre-and-post contexts to assist in the extraction process. There wrer common pre-and-post contexts such as the commonly recognized prepositions "for", "on", and "to".

7.1.5 Rules for Extraction of Cargo Information

In the case of cargo flights, the texts source reports the freight type in the active voice, using two patterns. The first pattern specifies the type of freight and its quantity, using weight measurement units such as kilograms, pounds, tons, etc., as in the following figure:



2011 The Antonov AN-12 cargo, operated by Trans Air Congo, took off from
Congo, bound to Pointe Noire, Congo. Four crewmember were onboard. The plane
le approaching Pointe Noire airport, killing all the 4 people onboard, and 19 people
he cargo plane had been largely empty, only carrying about 750 kg of beef. It
o a working-class residential area of Pointe-Noire port city as it
23 people (including the 4 crewmember and 19 fatalities on grou
ound, and razing around a dozen houses and causing a fire that
e aircraft's last seconds show it diving steeply and rolling to the
video appears to show exhaust trails coming only from the two
fatal roll was caused by the impossibility to counter with the ru
s on the same side. Pictures from the scene show the burnt-out
buildings in the suburbs of the city, with little remaining except

Features	
freight_type	beef
orth	mixed
quantity	750 kg
sem	cargo_type
syn	np
source	sarah.rul
rulid	cargo1

FIGURE 8: Extraction of cargo flight information**.**

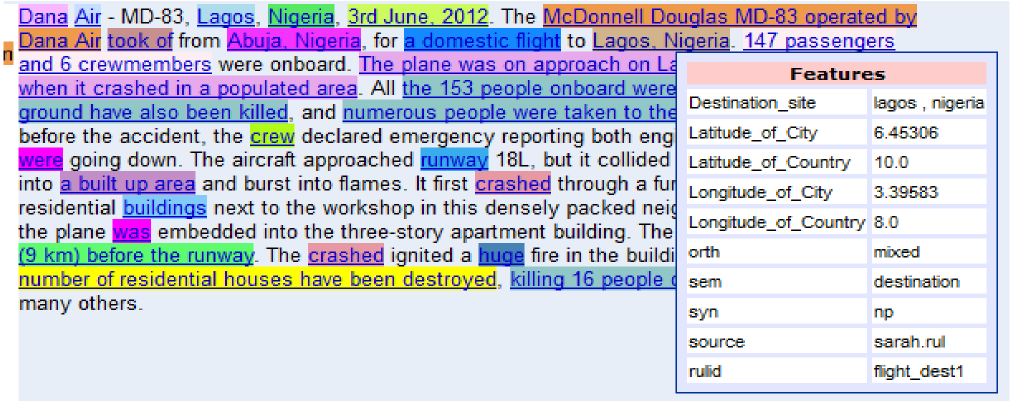
The above figure shows the annotated phrase under the semantic class “cargo_type” with the features “freight_type”, which is self-explanatory, and “quantity”, which defines the exact quantity of shipped cargo. Some of the measurement units have been found by default in the gazetteer semantic class "measure/exact_measure". New units can be added to the previously built-in class. The second pattern reports the type of cargo only, without defining its quantity, as in "carrying cars and various goods". Obviously, the main indicator of cargo information expression is the verbs such as "carrying", "loaded", and other that always precede the information about cargo. Those verbs and their possible tense variations were considered while designing the rules.

7.1.6 Rules for Extraction of Departure and Destination Sites

In 1001 crash, various formats are used to represent both departure and destination locations, separately. The general format includes the city, then the country as follows:

"took off from Odessa, Ukraine" (departure)
"to Donetsk, Ukraine" (destination)

The pre-and-post contexts of the extracted elements usually follow one common pattern in all the texts. This feature helps to distinguish between the departure and destination sites, or, later, even the crash site. In the case of departure, the regular pre-context is a combination of the verb phrase “took off” and the preposition “from”. With destination, the pre-context is a combination of the already-extracted semantic class “flight_type” and the preposition “to”.



Dana Air - MD-83, Lagos, Nigeria, 3rd June, 2012. The McDonnell Douglas MD-83 operated by
Dana Air took of from Abuja, Nigeria, for a domestic flight to Lagos, Nigeria. 147 passengers
and 6 crewmembers were onboard. The plane was on approach on La
when it crashed in a populated area. All the 153 people onboard were
ground have also been killed, and numerous people were taken to the
before the accident, the crew declared emergency reporting both eng
were going down. The aircraft approached runway 18L, but it collided
into a built up area and burst into flames. It first crashed through a fur
residential buildings next to the workshop in this densely packed neig
the plane was embedded into the three-story apartment building. The
(9 km) before the runway. The crashed ignited a huge fire in the buildi
number of residential houses have been destroyed, killing 16 people
many others.

Features	
Destination_site	lagos , nigeria
Latitude_of_City	6.45306
Latitude_of_Country	10.0
Longitude_of_City	3.39583
Longitude_of_Country	8.0
orth	mixed
sem	destination
syn	np
source	sarah.rul
rulid	flight_dest1

FIGURE 9: Extraction of destination information.

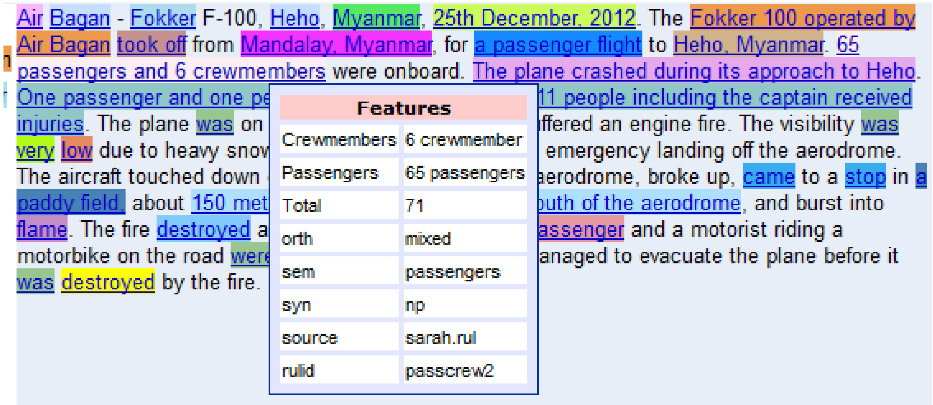
The city and country can be identified with assistance from the built-in gazetteer class entry "sem=geoname". The rules also benefited from the features provided with this class, such as latitude and longitude, to define the location accurately.

However, some problems have been encountered in extracting some cities and countries due to possible spelling diversity in translation locations names. Names like France and Odessa are both recognised as female personal names and are thus not extracted. This can happen if they were either not in the default gazetteer entries or written in different forms from their corresponding gazetteer entries. This problem can be addressed by entering the limited number of unrecognized countries to the gazetteer class "geoname" with their latitudes and longitudes. In terms of cities and states, the orthographic features such as "capitalised" had been used due to the large numbers of states and small cities that would need to be entered.

7.1.7 Rules for Passengers and Crew Numbers

The chosen source reports include the number of people on board at the time of the crash using various formats. The first format is the most commonly used, and is based on presenting the passenger numbers first, followed by crewmembers numbers, or vice versa. Some formats provide detailed information about the passengers, such as "37 passengers, 6 crew and 2 Sukhoi officials were on board". According to text patterns, different words are used to represent the airplane occupants, such as "passenger(s)", "crewmember(s)", "occupants", etc.

A new gazetteer class has been created to accommodate all these expressions under the name "plane_members". The following figure presents an example of this extraction:



Air Bagan - Fokker F-100, Heho, Myanmar, 25th December, 2012. The Fokker 100 operated by
Air Bagan took off from Mandalay, Myanmar, for a passenger flight to Heho, Myanmar. 65
passengers and 6 crewmembers were onboard. The plane crashed during its approach to Heho.
One passenger and one pe 11 people including the captain received
injuries. The plane was on iffered an engine fire. The visibility was
very low due to heavy snow emergency landing off the aerodrome.
The aircraft touched down aerodrome, broke up, came to a stop in a
paddy field, about 150 met outh of the aerodrome, and burst into
flame. The fire destroyed a assenger and a motorist riding a
motorbike on the road were anaged to evacuate the plane before it
was destroyed by the fire.

Features	
Crewmembers	6 crewmember
Passengers	65 passengers
Total	71
orth	mixed
sem	passengers
syn	np
source	sarah.rul
rulid	passcrew2

FIGURE 10: Flight-occupant information extraction.

The numbers of passengers or crew might be mentioned in two ways:

1- Numerically, as in “13 passengers”.
2- Textually, as in “two passengers”.

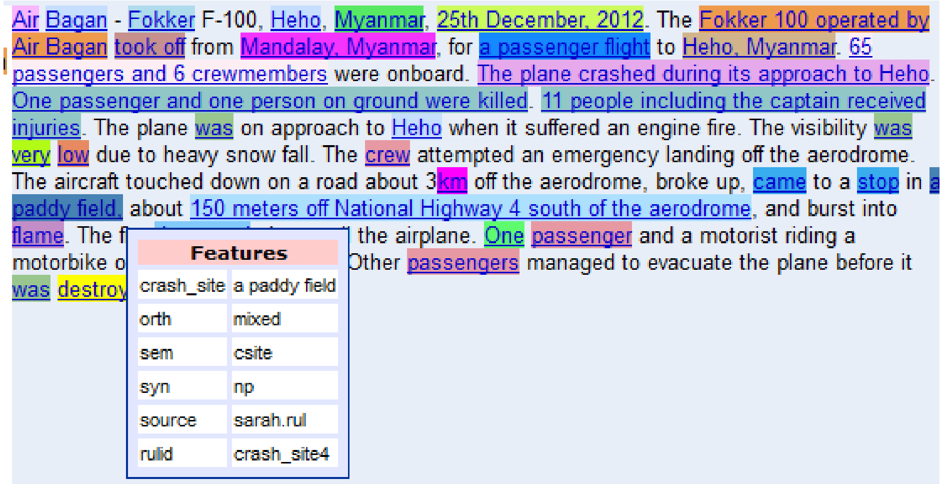
To cover the first pattern, the corresponding orthographic number feature has been used. In the second format, the default gazetteer class “measure/quantity/integer” is used to extract the textual format of the numbers. The unified post-context words “were onboard” have been considered for the purpose of efficiency. Additionally, a new attribute called “Total” has been added to some rules provide the total number of passengers and crew, in cases in which both are mentioned.

7.1.8 Rules for Crash Site Extraction

Crash sites were reported twice within the domain texts, but from different perspectives. In the first time, general sites are provided, such as mountains, lagoons, a city or village. In the second time, an accurate geographical dimension of the crash location is provided, which will be covered in the next rule.

According to the first mention of crash site, various locations of airplane crashes have been presented, such as mountains, fields, runways, etc. Two main patterns were discovered:

1- Some crash sites have been reported with the specific names of mountains, rivers, or cities, as in “The plane crashed on the bank of the Manohara river”. This can be extracted using the gazetteer default class “geoname” with the help of orthographic features. furthermore, the latitude and longitude features of the site can be found.
2- Some crash sites have been reported more generally, without accurate information about locations or names, such as in the following figure:



Air Bagan - Fokker F-100, Heho, Myanmar, 25th December, 2012. The Fokker 100 operated by Air Bagan took off from Mandalay, Myanmar, for a passenger flight to Heho, Myanmar. 65 passengers and 6 crewmembers were onboard. The plane crashed during its approach to Heho. One passenger and one person on ground were killed. 11 people including the captain received injuries. The plane was on approach to Heho when it suffered an engine fire. The visibility was very low due to heavy snow fall. The crew attempted an emergency landing off the aerodrome. The aircraft touched down on a road about 3km off the aerodrome, broke up, came to a stop in a paddy field, about 150 meters off National Highway 4 south of the aerodrome, and burst into flame. The f the airplane. One passenger and a motorist riding a motorbike o Other passengers managed to evacuate the plane before it was destroy

Features	
crash_site	a paddy field
orth	mixed
sem	csite
syn	np
source	sarah.rul
rulid	crash_site4

FIGURE 11: Crash site information extraction.

In this type, it was quite difficult to find a common pattern. The wide variety of possible places in which a plane may crash complicated the extraction. Therefore, this type of crash site was extracted by using the common orthographic features, as well as by defining the regular words number of the "crash site" phrase and its positions in text. As the proper name of the crash location was not mentioned, the latitude, longitude, and country features of those crash sites will not be found in the annotated token.

In terms of language structures, all texts defined the crash site in the active voice, starting with “the plane” and its synonyms, such as “aircraft”, “wreckage”, and others, which act as subject.

It was difficult to extract this type of entity because of multiple locations that were mentioned in the texts, including destination or other locations implicated in the crash, all of which reflect the need for specific trigger words.

The main trigger words are the verbs that report that a crash has happened. For example, in the example “The plane hit a lagoon called La Torrecilla“, the verb “hit” comes prior to the crash site information, which eases the extraction. This set of verbs and their different tenses have been added to the gazetteer under the semantic class “siteverb”.

There was an overlapping problem between extracting the crash site entity and the crash announcement event (which will be addressed later), due to the great similarity between their patterns and indicator verbs, such as “crashed”, which are used in both cases, as follows:

"The plane crashed on the bank of the Manohara river" (entity).
"The plane crashed during a go-around at Taladi airport" (event).

To avoid incorrect extraction, the sentence number and orthographic and syntactic features have been studied in-depth to be applied in the rules as restrictions in extracting crash site information. Another problem has been discovered with locations names that include hyphens and in which both parts are capitalised, such as “Petropavlovsk-Kamchatsky”. These are not compatible with the orth features “caphyphenated” or “lhyphenated" in CAFETIERE. This kind of case was extracted as three tokens using an extra rule.

7.1.9 Rules for Extraction of a Crash’s Geographical Dimensions

As clarified above, the crash site is reported again in the texts but with more details by specifying the incident’s distance from either one or two specific places.
Two main patterns for the crash dimensions are used within the domain texts, as in the following:

1- N "measurement_units" of “place”.
2- N "measurement_units" of “place 1”, N "measurement_units" of “place 2”.

The first pattern records the incident distance (N) according to one particular place, as in "The plane ended up some seven kilometres (four miles) from the runway". It is covered by a rule with two features: the position and distance of the affected plane which contain the values "from the runway" and "seven kilometers (four miles)", respectively.

The second pattern records the incident distance according to two different places, as in “about 200 meters off the coast and about 5 km north of the airport”. It is covered by a rule with two features: distance 1 and distance 2.

Both of these patterns were extracted with the assistance of their regular word sequence, and orthographic and syntactic features.

In terms of language structures, most crash dimension expressions are provided in the active voice, in which “plane” and its synonyms act as subject and are followed directly by the incident distance. In some cases, the subject is not followed directly by the crash distance, and this is considered when designing the rules. The main trigger words to distinguish this pattern from others are length measurement units such as “kilometres” and “feet”, and their abbreviations. All those units have been added to the gazetteer under the semantic class “measurement_unit”, taking into account the American spellings of the corresponding words.

Some texts reveal the crash distance without using measurement units such as: "The plane stopped on soft ground to the right of runway and parallel taxiway". This might lead to an overlapping problem with other, similar textual elements. This problem has been mitigated by using the pre-extracted element “crash_site” as pre-context and the new, added gazetteer class “crash_dim” as a constituent element. This class includes the common words of crash dimension patterns, such as “runway”, “taxiway”, etc.

The rule below annotates the textual phrase of the first pattern.

```
# Incident dimension from one place
[syn=np, sem=cdim, position=__a, distance=__b, type=entity, rulid=crash_dim6]=>
[sem=csite],
[token=","]?,
[orth=lowercase]*,
[token=","]?
\
[syn=CD, orth=number, token=__b],
[sem="measurement_units", token=__b],
[token="(", token=__b]?,
[syn=CD, orth=number, token=__b]?,
[sem="measurement_units", token=__b]?,
[token="(", token=__b]?,
[orth=lowercase|other|capitalized, token!=about, syn!=CC, token=__a]* /;
```

7.2. Rules for Extracting Relations

Relations can be used to answer questions like "Who operates this airplane?". Relations include two or more entities, depending on the patterns of the selected texts (Feldman & Sanger, 2007). Within the selected texts, one type of relation has been recognized; it appears below.

7.2.1 Belonging-to relation

The relation between the pre-extracted airplane type and its operator airline is the pattern targeted for extraction. In the belonging-to relation, the predefined semantic class “operateverb" is used to link the previous entities. This class includes all the phrases that are used in composing the belonging-to relation and is considered to be a keyword for it. However, in some texts the punctuation mark “,” is used additionally to join the two entities, which is counted as an option in the rules.

One general pattern of belonging-to relations is recognised in all the texts, such as: “The Dornier Do-228, operated by Sita Air”, and was covered successfully by the following rule:

```
#Belonging to relation
 [syn=np, sem=Belonging_to_relation, Airplane_type=_a, Airline_company=_b, type=relation,
rulid=belongto1] =>
\
[sem=airplane_information,token=_a],
[token=","]?,
[sem=operateverb],
[sem=airline_information,token=_b] /;
```

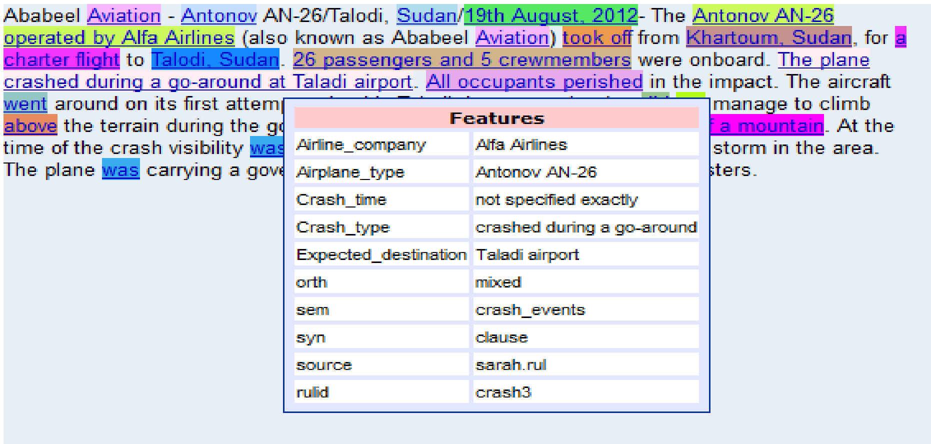
7.3. Rules for Extracting Events

Event extraction is considered the hardest element in the information extraction chain. A set of events have been recognized form texts and are listed below:

7.3.1 Rules for Crash Announcements

In the texts source, the formal declaration of the airplane crash is in the active voice. It begins with the determiner “the”, and then a word like “plane” or its synonyms, which are considered keywords. These announcements mainly include information about the type of crash that occurred, which means the circumstances to which the airplane was subjected. Usually, the scheduled destination is included in this phrase. Some texts include more detail by specifying the time of the crash. Consequently, the texts are grouped into three types, based on the details provided.

1. **The first type:** contains patterns in which the type of crash and, usually, the airplane's scheduled destination are mentioned. An example would be, "The plane crashed into the sea while attempting to land at Denpasar Airport". The following figure shows an example of this type of annotation:



Ababeel Aviation - Antonov AN-26/Talodi, Sudan/19th August, 2012- The Antonov AN-26
operated by Alfa Airlines (also known as Ababeel Aviation) took off from Khartoum, Sudan, for a
charter flight to Talodi, Sudan. 26 passengers and 5 crewmembers were onboard. The plane
crashed during a go-around at Taladi airport. All occupants perished in the impact. The aircraft
went around on its first attemp manage to climb
above the terrain during the go f a mountain. At the
time of the crash visibility was storm in the area.
The plane was carrying a gove sters.

Features	
Airline_company	Alfa Airlines
Airplane_type	Antonov AN-26
Crash_time	not specified exactly
Crash_type	crashed during a go-around
Expected_destination	Taladi airport
orth	mixed
sem	crash_events
syn	clause
source	sarah.rul
rulid	crash3

FIGURE 12: Crash announcement extraction, type 1.

2. **The second type:** includes phrases that announce the crash event in terms of crash type and crash time. Accordingly, the crash time has been taken from the texts that present it, using time measurement units, as in "The plane vanished from radar screens and lost contact with ground controllers after about 30 minutes of flight". The other text groups state the crash time in more general terms, such as "shortly after takeoff". Such unclear phrases are not considered to indicate crash times, due to the difficulty in finding a consolidated pattern among them. A default value has thus been added to the first and third text groups' rules: "crash_time=not specified exactly". This attribute value helps users recognise directly that these type of texts have not stated the time of the crash clearly.

3. **The third type:** covers the patterns that indicate the crash type, the crash site, and, sometimes, the scheduled destination, as in "The plane crashed in a residential area at Lake Kivu short of the runway while on approach to Goma's airport". In the first two text groups, the crash site was not mentioned in the same phrase of crash announcement, but later in the texts. Therefore, the crash sites in those two groups were extracted separately by another rule, which was discussed earlier in the section on entity rules.

As mentioned, the crash announcement event starts with the phrase "the plane" or the pronoun "it". Those phrases are considered "coreference" types because they refer in short form to the previously extracted element "airplane type and its operating airline", as seen below:
("The Antonov AN-24 operated by South Airlines" . . . . "The plane/It crashed")

Due to the CAFETIERE system's limitations in extracting coreferences, the coreferent cases cannot be referred directly to their main entities. Heuristics have been added to the rule contents to solve this problem. Extracting something in one sentence to relate it to something in another sentence is addressed by a temporary workaround on the beginning of rule structure as below:

```
[syn=clause, sem=crash_events, Airplane_type=_c, Airline_company=_d, crash_site=__a,
Crash_type=__b, crash_time= "not specified exactly", type=event,  rulid=crash4]
=>
```

```
[syn=np, sem=Belonging_to_relation, Airplane_type=_c, Airline_company=_d, type=relation,
rulid=belongto1],
[token!="WXYZ"]+,
[token="."]?
\
[orth=capitalized, token="The"|"It", sent>=2],
…….etc  /;
```

Obviously, the semantic class of the relation rule “belonging to” is used to define the airplane type and its operating airline, which is assigned as a pre-context. It is followed by the solution key which is [token !=”WXYZ"]+. This token is looking for anything that is not equal to the value “WXYZ” with the use of “+” or “*” iterators. Those iterators are greedy, and they match as much as they can. Thus, the developer has to choose words that definitely will never appear in the chosen texts. However, if the coreferent phrase appears repeatedly, then the system will pick only the last occurrence of this pattern, which is considered to be a drawback of this technique.

For example: “The Boeing 747-300….The plane….The plane…”.
The basic rule for extracting the entity of airplane type will extract “The Boeing 747-300”, and the last occurrence of “The plane” phrase.

In terms of trigger words, the regular pattern of a crash announcement includes a specific verb that confirms the crash. “Crashverb” is a new class that has been created in the gazetteer, and includes verbs such as “crashed”, “destroyed”, etc. Those verbs are mentioned in the texts in either the passive or the active voice which was considered while designing the rules.

Entities embedded within these events are also indicated by trigger words. First, the crash time is extracted by the already-extracted “crash_time” entity as part of the event rule. Second, the scheduled destination is generally mentioned after the prepositions “at” or "to", which are used as keywords. Third, the crash site may be located by looking for the preposition “at”. The "crash site" can be distinguished from the “scheduled destination” by its position in the clause. The "crash site" is usually mentioned before the scheduled destination and after the preposition “at”; this is taken into account when designing the rule. In some texts, the scheduled destination can be found after the punctuation mark “,” as a second clause in a crash event clause like, “The plane crashed after takeoff, close to its departure airport”.

7.3.2 Rules for Extracting Information about Casualties

In the chosen texts source, the crash casualties were divided into two levels, moderate and high, based on the details provided by the texts. Various patterns have described both categories.

In moderate-casualty reports, the number of surviving or evacuated people is mentioned alongside some injured cases without any killed cases. In high-casualty reports, the number of people killed, besides the injured cases and, sometimes, survivors are all mentioned.

These multiple pieces of information cannot be extracted as one event because they represent different facts about the crash they are describing. Therefore, in the first group, two types of rules were designed—one to extract the number of survivors, and a separate one to extract the number of injured. In the second group, three types of rules were designed to extract the number of killed, survived, and injured separately.

In the moderate-casualty situation, the survivors can be distinguished from survivors of high-casualty crashes by the capitalised determiner “All” which appears at the beginning of phrase as in, "All the people onboard survived". In high-casualty crashes, the number of survivors is indicated without using the determiner “All” as in, "47 people survived the accident". The semantic class of survivors in moderate-casualty crashes has been named “survivors”, while the one for high casualties has been named “people_affected”. In terms of injured persons, the pre-extracted "survivors" class is used as a pre-context to distinguish the injured people in moderate-casualty crashes from the ones in high-casualty crashes.

In terms of trigger words, many verbs were classified as indicators for both categories, such as “killed”, “injured”, or” “survived”. Those verbs are listed directly in the rules, as there is less variation in the action expressed by these verbs in the current texts source. Thus, the option of adding them to the gazetteer was skipped.

The examples below are of the moderate-casualty rule for injured persons:

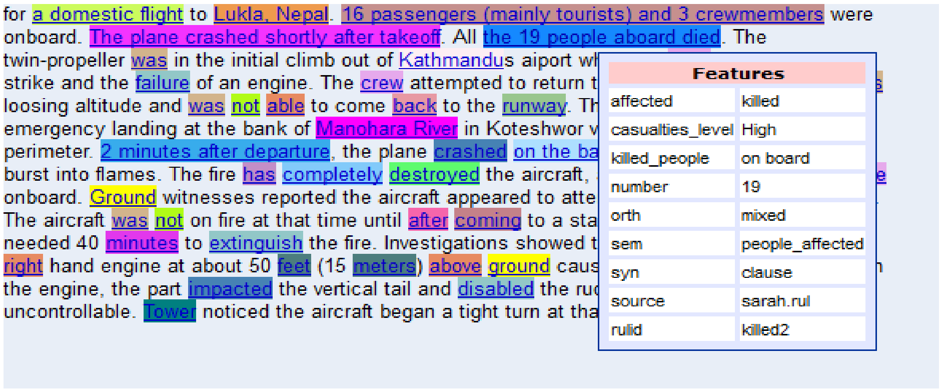
```
[syn=np, sem=people_affected, casualties_level=Moderate, affected=injured,
injured_people="on board", number=_n, type=event, rulid=injured5]=>
[sem=survivors],
[token= "."|","]?,
[syn=CC]?
\
[sem= "measure/quantity/integer", key=_n]?,
[syn=CD, orth=number, token~"^([0-9]{1,4})$",token=_n]?,
[sem="plane_members"]?,
[orth=lowercase]+,
[token="minor"|"serious"],
[token="injuries"] /;
```

In the high casualties group, the common patterns of survivors' phrases follow the same language structure as is used for survivors in moderate casualties, but without using the determiner “all”. In terms of people killed and injured, three groups have been created to classify them based on the casualty location: “on board”, “on ground”, and both, as follows:

- If all those killed were on board, then the texts mention the number of killed or injured people in general, without specifying the location, as in, “Five people were killed”.
- If there were people killed both on board and on the ground, then the texts mention the number of affected people and specify their locations, as in, “killing all seven occupants of the plane and as many as 25 people on the ground”.
- If those killed were all on the ground, then the texts mention the number of affected people, and specify that it happened on the ground, as in, “16 people were killed on the ground”.

The language structures from which this information was derived are divided into two patterns: active and passive voice.

The following figure shows the annotation results for people killed under the semantic class “people_affected” where the feature “casualties_level” has the value “High”; “affected” indicates that people have been killed; “number” specifies the number killed; and, lastly, “killed_people” defines the place of the people affected from the three potential values. The pre-defined gazetteer semantic class “plane_members” helps to specify the occupants’ type.



for a domestic flight to Lukla, Nepal. 16 passengers (mainly tourists) and 3 crewmembers were
onboard. The plane crashed shortly after takeoff. All the 19 people aboard died. The
twin-propeller was in the initial climb out of Kathmandus aiport wh
strike and the failure of an engine. The crew attempted to return t
loosing altitude and was not able to come back to the runway. Th
emergency landing at the bank of Manohara River in Koteshwor v
perimeter. 2 minutes after departure, the plane crashed on the ba
burst into flames. The fire has completely destroyed the aircraft,
onboard. Ground witnesses reported the aircraft appeared to atte
The aircraft was not on fire at that time until after coming to a sta
needed 40 minutes to extinguish the fire. Investigations showed t
right hand engine at about 50 feet (15 meters) above ground caus
the engine, the part impacted the vertical tail and disabled the ruc
uncontrollable. Tower noticed the aircraft began a tight turn at tha

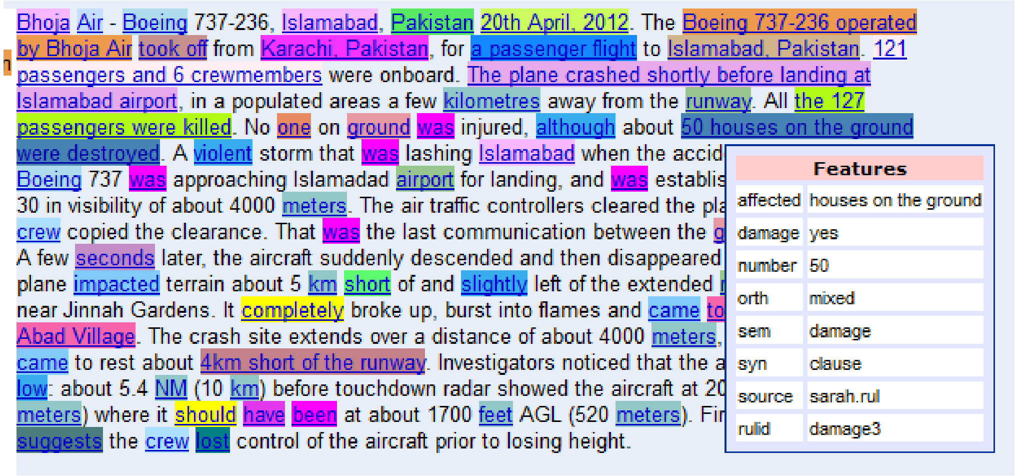
Features	
affected	killed
casualties_level	High
killed_people	on board
number	19
orth	mixed
sem	people_affected
syn	clause
source	sarah.rul
rulid	killed2

FIGURE 13: Extraction of information about people killed.

7.3.3 Rules for Extracting Damage

Property damage was reported as a consequence of some air incidents covered in the studied texts, as in "50 houses on the ground were destroyed". The types of affected property are described in the texts as houses, buildings, dwellings, etc.

As such, "properties" and "damageverb" were created as new gazetteer classes to accommodate the various building types and regular damage verbs as entries, respectively. Those entries were considered as main indicators of damage information expression. In term of language structures, there were only two patterns detected in reporting damage: either the active or the passive voice which were considered during rules design. The following figure presents an example of extraction of damage information passively:



Bhoja Air - Boeing 737-236, Islamabad, Pakistan 20th April, 2012. The Boeing 737-236 operated
by Bhoja Air took off from Karachi, Pakistan, for a passenger flight to Islamabad, Pakistan. 121
passengers and 6 crewmembers were onboard. The plane crashed shortly before landing at
Islamabad airport, in a populated areas a few kilometres away from the runway. All the 127
passengers were killed. No one on ground was injured, although about 50 houses on the ground
were destroyed. A violent storm that was lashing Islamabad when the accid
Boeing 737 was approaching Islamadad airport for landing, and was establis
30 in visibility of about 4000 meters. The air traffic controllers cleared the pla
crew copied the clearance. That was the last communication between the g
A few seconds later, the aircraft suddenly descended and then disappeared
plane impacted terrain about 5 km short of and slightly left of the extended r
near Jinnah Gardens. It completely broke up, burst into flames and came to
Abad Village. The crash site extends over a distance of about 4000 meters,
came to rest about 4km short of the runway. Investigators noticed that the a
low: about 5.4 NM (10 km) before touchdown radar showed the aircraft at 20
meters) where it should have been at about 1700 feet AGL (520 meters). Fir
suggests the crew lost control of the aircraft prior to losing height.

Features	
affected	houses on the ground
damage	yes
number	50
orth	mixed
sem	damage
syn	clause
source	sarah.rul
rulid	damage3

FIGURE 14: Extraction of damage information.

Additionally, rules can extract either the past or present perfect verb tenses using the optional character “?”, with the built-in gazetteer classes "beverb", and "haveverb". Some texts report on damage generally, without specifying an exact number of affected buildings which have covered by rules without the attribute "number".

8. DISCUSSION

The implementation and testing phases encountered difficulty when it came to the development of sometimes-complicated extraction rules whose complexity was underestimated in the initial design phase. In the case of entity recognition, the date, airplane type, airline, passenger numbers, and crash site were estimated as straightforward elements to be easily extracted due to their clear and consistent patterns. However, language structures for airplane type and crash site have stepped out from this set. It was expected that the airplane type would start with the capitalized manufacturer name followed by numbers, to be extracted as only two tokens, such as “Sukhoi 100”. However, some instances of airplane type did not conform to this simple pattern, and were extracted as multiple tokens, especially when they contained a set of symbols separated by hyphens. For example, “Lockheed C-130J-30 Hercules” does not match the orthographic features “caphyphenated” or “lhyphenated”, resulting in a need to treat them as eight tokens extracted using more complicated rules.

The complexity of the rules around crash sites was also underestimated. A set of obstacles was faced in writing these rules. First was the huge possible range of crash sites and their different geographic characteristics. Those locations were mentioned either with or without names. For those without specific names, the orthographic and the syntactic features, as well as the regular number of words in each phrase, were used as strong clues. For those with specific names, the crash sites were covered with the help of the default gazetteer class “geoname”. Another problem arose in particularly long reports due to the multiple locations that were mentioned in the text as having possibly been affected by the plane crashing. Those location details complicated the process of designing the crash site rules. This led to consideration not only of the orthographic and syntactic features, but also of the regular position of the phrase within the text body using the sentence number feature “sent”.

On the other hand, the opposite situation was observed for the relation rule. It was classified as a challenging extraction whose ultimate ease surpassed expectations due to the fact that it depended on highly reliable extracted entities as well as limited patterns, which facilitated the task. In terms of events, the projected difficulty emerged as expected. A wide range of patterns within a single event was discovered. This situation complicated the process of designing efficient event rules.

In terms of events structure, there are two types of events: dependent and independent. The dependent events must be based on fully successful extractions of participant entities and relations. Formulating rules for dependent events is considered a reasonably hard task. In the case of independent events, they do not depend on previously extracted elements, and the extraction was thus quite a bit easier. The general strategy that was adopted for all event cases depended on extracting each event separately, which eased the rule-writing process considerably.

CAFETIERE’s formalism helps to reduce the complexity of events further. For example, it allows different verb tenses to be grouped in one rule to cover all the possible tense variations easily.

Within the chosen text source, some spelling mistakes were observed during analysis which might have an impact on some rule implementation. The words spelling mistakes could not be rectified, and only affected system workflow if they were trigger words. As such, these mistakes have been considered while designing the rules. In addition, the texts source uses the American English style which has been considered in the words spelling during the rule design. For example, the words “kilometer”, ‘meter” and other measurement units have been entered to the gazetteer to be recognised easily.

In terms of long reports, the dimensions of secondary crash sites that a crashing plane might hit before coming to a rest were also considered. Those sites used phrases similar to those used in the main crash dimensions. They used measurement units, such as kilometres and feet, which were considered the main indicators for crash dimension patterns. This issue led to some cases of spurious extraction of crash dimensions. However, this problem was addressed to some extent by the use of the crash site as a pre-context.

Even though the most important features of domain texts have been studied, further improvements might have been obtained if there had been sufficient time. Reducing the number of partial and spurious extractions was at the top of the list of desired improvements to be effected by further analysis.

9. SYSTEM EVALUATION

This chapter describes the adopted evaluation method for the developed information extraction rules. The performance of the system is evaluated using MUC measures. The results are analysed with the use of tables.

9.1. Evaluation Measures

The evaluation stage represents an important part of any project development chain due to its role in measuring the quality of finished work. As mentioned earlier, in chapter two, MUC provided common evaluation metrics, which are precision (P), recall (R), and F-measure.

As explained, recall calculates the accuracy of the system, while precision measures the coverage level. The F-measure provides an overall score for system performance.

In the evaluation process, a set of texts was annotated manually to be compared with the same set of texts that were annotated by the IE system. The manual annotations represent the elements that the developer expects the system to extract. The results of comparing the two types of annotations can be categorised as follow:

- Correctly extracted: relevant information that has been annotated correctly in both manual and system annotations.
- Spurious: elements that have been annotated by the system and not annotated manually.
- Missing: elements that have been annotated only manually and have not been annotated by the system.
- Incorrectly extracted: elements that have been annotated by the system incorrectly.

Additionally, in some results, only part of the elements was extracted by the system. Those were considered to be partially extracted elements. Thus, a coefficient must be present to score the incompleteness of the result, which is generally equal to a 0.5 coefficient (Turmo et al., 2006). An example of partial extraction is a case in which the element to be extracted is an airline, such as "National Air Cargo Services", but the result of extraction is “National Air”. In this situation, the 0.5 coefficient is used. To widen the scope of accuracy, more values can be assigned to the coefficient. For example, 0.75 and greater can be assigned if the major part of the element is extracted, which reflects higher accuracy. Conversely, 0.30 and lower can be assigned if only a small part of the element was extracted, which represents lower accuracy.

In fact, the new edited formula for recall, which considers partial extraction, is given below:

$$R = \frac{N_{correct} + 0.5N_{partial}}{N_{manual}}.$$

Also, the formula for precision has been modified to consider partial extraction as follows:

$$P = \frac{N_{correct} + 0.5N_{partial}}{N_{correct} + N_{incorrect} + N_{spurious}},$$

It is worth noting that the coefficient value in the above formula is assigned to be 0.5. It is possible to change its value under the previously mentioned conditions.

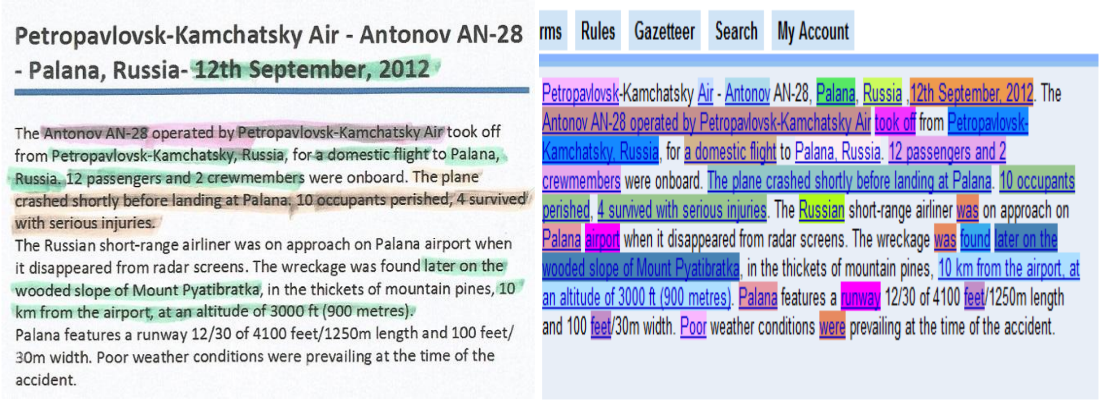
9.2. Evaluation Process

The project adopts the same evaluation methodology that was used in MUC. The pre-defined evaluation metrics were calculated first based on the sample texts that were used during the process of rule design. The high degree of accuracy in the results here might be due to the use of these same texts for both evaluation and rule development. For the sake of efficiency, a set of randomly selected texts taken from the same website was used as test texts in order to measure the real performance of the system. The texts, of varying lengths, all pertain to events happening between 2011 and 2013.

There were no guidelines to be followed in evaluating the CAFETIERE systems. Because of this, a new strategy was adopted to regulate the calculations process for extracted elements; this strategy is outlined below.

- Only the elements extracted according to the developed rules will be considered during the evaluation. The main objective of this project is to measure the quality of extraction rules when extracting the required information.
- Five main categories have been created to enhance the calculation of extracted entities, relations, and events. Those categories are “correct”, “partial”, “missing”, “incorrect”, and “spurious” elements.
- The criteria for assigning the values of coefficients for partially extracted elements must be clarified early. A decision was thus made that the coefficient values would range from 0.30 to 0.50 to 0.75, which means that the extracted element might fewer than half match, half match, or more than half match, respectively.
- The calculation for P, R and F measures have been done separately for entities, relations, and events in order to achieve accurate results.
- The built-in and domain gazetteer entries are not counted in the calculation process in this project. This decision was made because of the system’s 100% accuracy in highlighting the gazetteer entries in all texts. Besides, most semantic classes within a gazetteer have been used in designing the rules themselves, so that their mutual influence is evident.

The following figure presents a comparison between the manual annotations and the system annotations of the same text, which has been applied to both test and working texts,:



Petropavlovsk-Kamchatsky Air - Antonov AN-28 - Palana, Russia- 12th September, 2012

The Antonov AN-28 operated by Petropavlovsk-Kamchatsky Air took off from Petropavlovsk-Kamchatsky, Russia, for a domestic flight to Palana, Russia. 12 passengers and 2 crewmembers were onboard. The plane crashed shortly before landing at Palana. 10 occupants perished, 4 survived with serious injuries.

The Russian short-range airliner was on approach on Palana airport when it disappeared from radar screens. The wreckage was found later on the wooded slope of Mount Pyatibratka, in the thickets of mountain pines, 10 km from the airport, at an altitude of 3000 ft (900 metres).

Palana features a runway 12/30 of 4100 feet/1250m length and 100 feet/30m width. Poor weather conditions were prevailing at the time of the accident.

Petropavlovsk-Kamchatsky Air - Antonov AN-28, Palana, Russia ,12th September, 2012. The Antonov AN-28 operated by Petropavlovsk-Kamchatsky Air took off from Petropavlovsk-Kamchatsky, Russia, for a domestic flight to Palana, Russia. 12 passengers and 2 crewmembers were onboard. The plane crashed shortly before landing at Palana. 10 occupants perished, 4 survived with serious injuries. The Russian short-range airliner was on approach on Palana airport when it disappeared from radar screens. The wreckage was found later on the wooded slope of Mount Pyatibratka, in the thickets of mountain pines, 10 km from the airport, at an altitude of 3000 ft (900 metres). Palana features a runway 12/30 of 4100 feet/1250m length and 100 feet/30m width. Poor weather conditions were prevailing at the time of the accident.

FIGURE 15: Comparison between system and manual annotations.

As shown, the manual annotations appear on the left side of the screen, while the system annotations appear on the right side. The entities, relations and events are highlighted in green, pink, and orange respectively. The complete match between the manual and system annotations is evident.

9.3. Analysis of Metrics for Training Data

The table below shows the calculation of those elements extracted systematically and manually for the training texts. The correct, spurious, missing, partial and incorrect annotated elements extracted by the system are counted.

Texts	System Annotations					Manual Annotation
	Total	Correct	Partial	Spurious	Incorrect	
1	9	9				9
2	9	9				9
3	11	9	1	1		10
4	10	10				10
5	9	9				9
6	10	8	2			10
7L	10	9		1		11
8	9	9				9
9	9	9				9
10	9	9				9
11	9	9				9
12L	10	10				10
13	6	6				6
14	8	8				8
15	9	9				9
16	10	10				10
17	9	7	2			9
18	9	9				9
19	8	7	1			9
20	9	9				9

TABLE 3: Training data results for entities.

Texts	**Relations Extracted**				**Manual Annotation**	**Events Extracted**				**Manual Annota -tion**
	Correct	Partial	Spurious	Total		Correct	Partial	Spurious	Total	
1	1			1	1	6			6	6
2	1			1	1	2			2	2
3	1			1	1	2			2	2
4	1			1	1	2			2	2
5	1			1	1	3			3	3
6	1			1	1	3			3	3
7L	1			1	1	3			3	3
8	1			1	1	5			5	5
9	1			1	1	3	1		4	4
10	1			1	1	1	1		2	2
11	1			1	1	3			3	3
12L	1			1	1	5			5	5
13	1			1	1	3			3	3
14	1			1	1	2			2	2
15	1			1	1	3			3	3
16	1			1	1	6			6	6
17	1			1	1	3			3	3
18	1			1	1	3			3	3
19	1			1	1	2			2	2
20	1			1	1	7			7	7

TABLE 4: Training data results for relations and events.

The precision, recall, and F-measure were calculated for the previously tabulated values using the redefined formulae, which are listed below.

Texts	**Entities**			**Relations**			**Events**		
	Precision	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	0.89	0.97	0.93	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	0.90	0.90	0.90	1.00	1.00	1.00	1.00	1.00	1.00
7L	0.90	0.81	0.85	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.94	0.94
10	1.00	1.00	1.00	1.00	1.00	1.00	0.87	0.87	0.87
11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
12L	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
17	0.94	0.94	0.94	1.00	1.00	1.00	1.00	1.00	1.00
18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
19	0.94	0.83	0.88	1.00	1.00	1.00	1.00	1.00	1.00

20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

TABLE 5: Evaluation measures values for the training data.

The tabulated values reflect the system's performance. In terms of entities, the results are not as accurate as those for events and relations, but they are still considered highly accurate. The degree of accuracy of results ranges from 85% to 100% in F-measure, because of the iterative approach used in designing the rules. However, there were some challenging entities to be extracted. For example, the flight purpose might be presented differently in some texts, with details that were hard to assess in terms of flight purpose such as the following example: "Sukhoi was performing a demonstration flight for airline representatives and journalists in order to promote sales of the aircraft". The more common pattern was always a short form, such as "a passenger flight". This situation led to partial extraction, and thereby decreased the precision scores.

In terms of relations extraction, tabulated values reflect the tremendous performance of the system compared to its performance with entities. All the defined relations in the working texts were successfully extracted. Precision, recall and F-measure values were equal to 1, which reflects a high degree of accuracy in recognising relations. In terms of events extraction, the system also shows accurate results with most texts, with precision equal to 1. In two cases, the precision and recall values are lower than 1, which means that some events have been annotated partly or spuriously. However, the overall system scores still show high-quality performance in extracting events. This is because the vast majority of event rules are independent. Only few types of event rules depend on previously extracted elements such as crash announcement extraction. Those rules might be affected by some failed results for entities extraction, leading to a possible reduction in the scores for events evaluation. Despite the fact that dependent events were based on reliable entities, their extraction was quite difficult compared to that of independent events.

9.4. Analysis of Metrics for Test Data

The same process of evaluating the working text has been followed for the test set. The tables below show the results of 10 randomly selected reports that were analysed.

Texts	Entities extracted					Manual Annotation
	Total	Correct	Partial	Spurious	Incorrect	
1	9	9				9
2	10	10				10
3L	8	5	1	2		6
4	9	9				10
5	9	9				9
6	8	7			1	7
7	9	9				10
8	8	8				8
9	9	9				9
10	10	10				10

TABLE 6: Test data results for entities.

Texts	Relations Extracted				Manual Annotation	Events Extracted				Manual Annotation
	Correct	Partial	Spurious	Total		Correct	Partial	Spurious	Total	
1	1			1	1	3			3	3
2	1			1	1	2			2	2
3L	1			1	1	2			2	2
4	1			1	1	2	1		3	3
5		1		1	1	3	1		4	5
6	1			1	1	1			1	1
7	1			1	1	2	1	2	5	3
8		1		1	1	4			4	4
9	1			1	1	6			6	6
10	1			1	1	2			2	2

TABLE 7: Test data results for relations and events.

In both tables, the results for the test reports are somewhat lower when compared to the sample corpus. This is caused by the new patterns involved in the test set, which had not previously been encountered within the rules.

In the test set, the system’s performance in extracting entities and relations is better than for events. This is because of new event patterns that were not covered in the process of rule development. In terms of relations and entities, the system performed very well, with just few unsuccessful extractions. For example, the relation between airplane type and airline might be annotated partly. This could happen if the airline companies or airplane types have been written in forms differ slightly from those that were considered during rules development. When it came to entities, the extraction of crash site information was quite challenging, due to the multiple minor locations that are embedded with the body of the long texts which might raise the number of spurious cases. However, the system evaluation metrics still showed high performance, evident in the tables below.

Texts	**Entities**			**Relations**			**Events**		
	Precision	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3L	0.72	0.96	0.82	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	0.90	0.95	1.00	1.00	1.00	0.83	0.83	0.83
5	1.00	1.00	1.00	0.75	0.75	0.75	0.94	0.75	0.83
6	0.87	1.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00
7	1.00	0.90	0.95	1.00	1.00	1.00	0.50	0.83	0.62
8	1.00	1.00	1.00	0.75	0.75	0.75	1.00	1.00	1.00
9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

TABLE 8: Evaluation measures values for the test data.

As shown, the entities extraction reflects very good performance, with success rates of 82% and higher. This proves the efficiency of the developed rules. Relations also achieved high scores of mostly 100%, with only two exceptions. The reason behind the high success rate with relations is the limited patterns present in the texts. The lower scores cases were the result of relations' dependency on some unsuccessfully extracted entities, which was explained earlier. Greater fluctuations were seen in the results for events, whose success rates range from 62% to 100%. This is different from event results in the working texts. The reason for this is the complexity of and differences among event patterns, which were new and difficult to cover even with generic rules.

The main observations in both result groups are as follows. In the working texts, the performance level for relations and events extraction was higher than that for entities. In the test texts, event performance showed slightly lower success rates compared to those of entities and relations, which were explained earlier. Moreover, the results for the long reports, which are marked with a capitalised L in the tables, are less efficient than the results for the short ones. As explained previously, the long reports include undesirable minor details that have phrase structures similar to the structures of main patterns to be extracted. This might lead to unsuccessful extractions such as the previous explained case of crash site.

In order to map the performance for individual entities and events in both the test and training texts, the table appearing below was created. Relations have not been counted due to the similarity of the results for both sets. Table 9 shows the breakdown of the results. Some entities and events show extremely high performance, including airlines, departure and destination (entities) alongside casualties and damage (events).

Entities	Training Texts Extractions					Test Texts Extractions				
	Correct	Partial	Spurious	Missed	Manual Annotation	Correct	Partial	Spurious	Missed	Manual Annotation
Crash Time	2				2	2				2
Crash Date	20				20	10				10
Airplane Type	19			1	20	8	1		1	10
Airline Company	20				20	10				10
Departure Site	20				20	10				10
Destination Site	20				20	10				10
Flight Purpose	18	1	1		19	10				10
Passengers Number	20				20	10				10
Crash Site	18	5	2		24	8	1	1		10
Crash Dimensions	24				24	10				10
Cargo Type	5		1		6	2				2
Events										
Crash announcement	20	2			22	6	2		2	10
Casualties	44				44	19	1			20
Damages	6				6	3	1			4

TABLE 9: An outline for entities and events results.

It is obvious that "airplane type" and "crash site", along with "crash announcement", have the greatest number of missed, spuriously or partially extracted elements. These results tally with the previous discussed reasons for this poor performance. For example, in the case of a crash announcement event, there is a problem with the use of some verbs, such as "crashed" and similar verbs, that are also used when relaying information about crash sites. This problem was addressed by using new classes specified for crash site and announcements verbs separately. The use of those classes along with in-depth study of both crash site and announcement regular patterns, besides the use of the "sent" feature mitigate unsuccessful extractions.

To sum up, it is safe to say that the performance of entity, relation, and event extractions was very good, even when taking into consideration time limitations and the developer's low level of experience in designing the rules. Additionally, the extraction results of the current project are highly comparable with MUC or ACE performance in newswires extraction results. For example, the state of the art for event extraction from newswires is around 60% which is successfully achieved.

9.5. Comparative evaluation:

According to MUC-7 evaluation results, the top scoring IE systems for extracting airplane crashes information from online newswire had achieved results as follows [22][23]:

- 95% for named entity recognition task.
- 60-80% for co-reference.
- 70-85% for relations, and
- 50-70% for events.

However, it is worth noting that a direct comparison between precision, recall and F-measure obtained by IE systems in various challenges is complex and indirect due to a set of difference factors such as: different template slots structure as well as the quality and the language of data set to be processed.

Consequently, a comparative evaluation has been conducted. It shows better results of the proposed system performance compared to results obtained by another system called ESSENCE which is also evaluated using the MUC scoring system. The results of the test corpus in both the proposed system and ESSENCE system are present in the below table. The table shows the precision, recall and F-measure metrics values for named entity (NE) task. As previously mentioned, the proposed system is a rule-based system in contrast to ESSENCE which is built with the use of a machine learning algorithm called ELA [22].

Entity	**Proposed System**			**ESSENCE**		
	Precision	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
Crash site	72.7	80.0	76.1	51.2	59.4	55.0
Crash Date	100	100	100	82.6	75.4	78.8
Aircraft	88.8	80.0	84.1	100	65.0	78.8
Airline	100	70.4	82.6	55.2	65.2	59.9
Departure	100	100	100	51.5	57.6	54.4
Destination	100	100	100	72.9	60.7	66.3
Average	93.5	88.4	90.4	68.9	63.8	65.5

Table (10): Comparative evaluation for NE task between ESSENCE and the proposed system.

Results presented show an average level of 65.5% in ESSENCE and 90.4% in the proposed system that is considered a quite high performance compared with the ESSENCE system in same tasks. The well-designed rules, which are based on an intensive study of text patterns, are the reason behind this achievement in the proposed system [22].

10. CONCLUSION

Developments within the field of information extraction have enhanced performance related to querying, regulating, and analysing data via the application of effective techniques for storing unstructured texts in structured formats. This project sought to gain a deep understanding of the information extraction field by designing a rule-based IE system for airplane crash reports that would extract useful information from a set of texts.

Reaching a sufficient level of understanding the information extraction field involves covering some aspects as follows. The definition of information extraction in terms of project workflow was given. The framework for evaluating information extraction systems, along with the overall extraction process, was provided.

In terms of the proposed system, a considerable investigation into information extraction domains was conducted, leading to the selection of the area of airplane crashes for the writing of rules according to a knowledge engineering approach. The text sources for this domain were carefully chosen to ensure accessibility, abundance, and the recognition of regular patterns. The formalism of the CAFERIERE web-based information extraction system and its components were studied in-depth. According to the methodology of this project, rules were designed, implemented, and then tested iteratively.

In terms of the selected corpus, a significant analysis was conducted in order to find common patterns within those reports.

The rules for this corpus were designed to enfold a combination of waterfall and prototype methods, taking into account the rules order conditions. The efficiency of rules performance was also checked throughout by calculating the common precision, recall, and F-measure metrics for all extracted elements. The results of the evaluation were discussed, and reflected the very high performance of the developed rules. The values of those metrics are at times equal to 1.0, even for the set of test texts, which is a result comparable to the scores of the MUC program.

Efficient system workflow can be accomplished in a number of ways, such as by choosing domain source texts that provide similar and consistent language patterns in both working and test corpora. Those contributory factors can lead to good-quality, well-developed rules in a short period of time. However, more efficacious rules with more accurate results might emerge if more time were spent on examining a greater number of patterns for the domain texts. This could be a goal in future work.

11. FUTURE WORK

In this paper, an IE system for airplane crash reports have been presented using the rule-based approach. The main goals have been achieved successfully through an intensive works using an effective methodology. The advantage of this methodology is that it reduces the chances of wrong extraction results in the process of developing an IE system.

However, due to time restrictions, only critical elements were extracted in this study. The system could, be broadened in the future to include extraction of reasons for crashes as well, involving such things as weather, technical malfunctions, or terrorism. These could be stored in databases or on the web for further global aviation analysis using data-mining techniques.

The selected corpora might be widened to include more cases of air incidents, leading to more patterns and more comprehensive analysis, and thereby to higher-performance systems. A system could also be developed to be part of a larger application that, for example, supports web research into air incidents. To achieve this, several issues first need to be addressed. First, many location patterns such as departures, destinations, primary crash sites, and the final location of the affected plane are often mentioned in a single report. Those need to be covered accurately with further analysis of an extra set of texts to cover all possible patterns. Second, within this project it was presumed that the only airplane type information to be extracted was that mentioned in the body of a text. Other formats for airplane type, such as those appearing in a text’s title, were skipped. The type was usually mentioned in the title in more detail, and generally did not change the extraction facts; to raise the level of accuracy, however, it would be preferable to differentiate between the details within the different formats. Third, as mentioned earlier, the supporting of co-reference resolution will significantly enhance the efficiency of the developed system. Finally, the implementation of a spell-checker for rules will accelerate the process of development by eliminating time lost in searching for spelling mistakes in extraction rules.

12. REFERENCES

[1] Appelt, D.E. (1999) "Introduction to information extraction", [Online] Available from: http://philarts.spbu.ru/Members/lida_pivovarova/Appelt.pdf. [Accessed on: 3/03/2016].

[2] Appelt, D. and Israel, D. (1999) "Introduction to Information Extraction Technology: IJCAI-99 tutorial', [Online] Available from: http://www.ai.sri.com/~appelt/ie-tutorial/IJCAI99.pdf [Accessed on 28/05/2016].

[3] Ben-Dov, M. and Feldman, R. (2005) "Text Mining and Information Extraction". In: Maimon, O. and Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook. Springer Science + Business Media, Inc., pp. 801-831.

[4] Black, W.J., McNaught, J., Vasilakopoulos, A., Zervanou, K., Theodoulidis, B., and Rinaldi, F. (2005) "CAFETIERE: Conceptual Annotations for Facts, Events, Terms, Individual Entities, and RElations", Parmenides Technical Report TR-U4.3.1, [Online] Available from: http://www.nactem.ac.uk/files/phatfile/cafetiere-report.pdf [Accessed on 13/04/2016].

[5] Black, B. (2007) "Cafetiere Users' Guide". [Accessed on 13/04/2016].

[6] Black, B. (2013) "Cafetiere Users' Guide". [Accessed on 20/06/2016].

[7] Chinchor, N. (1992) "MUC-4 Evaluation Metrics", [Online] Available from: http://acl.ldc.upenn.edu /M/M92/M92-1002.pdf. [Accessed: 13/06/2016].

[8] CMS (2008) "Selecting a Development Approach", [Online] Available from: http://www.cms.gov/ Research-Statistics-Data-and-Systems/CMS-Information-Technology /XLC/Downloads/SelectingDevelopmentApproach.pdf. [Accessed: 8/04/2016].

[9] Cowie, J. and Lehnert, W. (1996) "Information Extraction", Communication of the ACM, 39(1), pp. 80-91.

[10] Feldman, R. and Sanger, J. (2007) The Text Mining Handbook: Advanced Approaches In Analyzing Unstructured Data. New York: Cambridge University Press.

[11] Grishman, R. (1997) "Information Extraction: Techniques and Challenges." In: Pazienza, M.T. (ed.) Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology. Berlin, Heidelberg: Springer-Verlag, pp. 10-27.

[12] Jones K.S. and Gallier J.R. (1996) "Evaluating Natural Language Processing Systems: An Analysis and Review", Springer, Berlin.

[13] Kao, A. and Poteet, S.R. (Eds.). (2006)"Natural Language Processing and Text Mining" London, UK: Springer-Verlag, pp. 12-40.

[14] Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E., and Soderland, S. (1994) "Evaluating an Information Extraction System," Journal of Integrated Computer-Aided Engineering, 1(6).

[15] Lewis, David D. (1995). Evaluating and Optimizing Autonomous Text Classification Systems. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM, pp. 246-254.

[16] McDonald, D., Kelly, U., McNicoll, L. and Weir, G. (2012) "The Value and Benefit of Text Mining to UK Further and Higher Education. Digital Infrastructure." [Online] Available from: http://bit.ly/jisc -textm. [Accessed: 01/03/2016].

[17] Redfearn, J., JISC Communications team. and Nactem. (2006), “Text Mining,” [Online] Available from: http://www.nactem.ac.uk/files/papers/JISC-BP-TextMining-v1-final.pdf. [Accessed: 28/02/2016].

[18] Riloff, E. and Lorenzen, J. (1998) “Extraction-Based Text Categorization: Generating Domain-Specific Role Relationships Automatically” [Online] Available from: http://www.cs.utah.edu/~riloff/ pdfs/nlp-ir-chapter.pdf. [Accessed: 2/03/2016].

[19] Sitter, A.D., Caldersy, T., and Daelemans, W. (2004) “A Formal Framework for Evaluation of Information Extraction” [Online] Available from: http://wwwis.win.tue.nl/~tcalders/pubs /DESITTERTR04.pdf. [Accessed: 17/04/2016].

[20] Turmo, J., Ageno, A., and Catala, N. (2006) “Adaptive Information Extraction,” ACM Computing Surveys, 38(2), pp. 1-47.

[21] Zhong, N., Li, Y., and Wu, S.T. (2012) “Effective Pattern Discovery for Text Mining,” IEEE Transaction on Knowledge and Data Engineering, 24(1), pp. 30-44.

[22] Català, N., Castell, N., and Martín, M. 2000. ESSENCE: A portable methodology for acquiring information extraction patterns. In Proceedings of the 14th European Conference on Artificial Intelligence (ECAI), pp.411-415. [Online] Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.207&rep=rep1&type=pdf [Accessed: 28/01/2017].

[23] Poibeau,T., Saggion, H., Piskorski, J. and Yangarber, R. (2013) ”Information Extraction: Past, Present and Future” in Multi-source, Multilingual Information Extraction and Summarization, 1st ed. Berlin: Springer- Verlag Berlin Heidelberg, pp.23-49. [Online] Available from: http://www.springer.com/cda/content/d...1.pdf?SGWID=0-0-45-1342906-p174307561. [Accessed: 2/02/2017].

[24] Alkadi, S. “Information Extraction.” Master thesis, University of Manchester, U.K., 2013.