A NEW APPROACH FOR MULTILINGUAL EVENTS EXTRACTION (OraMEE)

ABDEL ALNASSER A. ALASFOUR [1], ŞTEFAN TRĂUŞAN-MATU[2]

Internet has become a global information hub, covering almost every aspect of life in the form of books, pictures, videos etc. in many different languages. This information is growing rapidly, and is increasingly becoming available to everyone. However, it is very difficult for a user to take advantage of the information available in languages not known to him/her. In order to overcome the language limitations, many ideas have been surfaced, e.g. Cross Languages Information Retrieval (CLIR) and Multilingual Information Extraction (MLIE) etc. Now, with the ease of internet access, any events occurring anywhere in the world are also spread instantly over the internet. These events may emerge as standalone events or as part of a multi-event hierarchy, and can be spread in any language. In this article, we will present an approach OraMEE for automatic event extraction from multilingual text.

Keywords: CLIR, Oracle, Multilingual Event Extraction, MLIE, OraMEE, IR, IE, Gist, Themes

1. Introduction

With rapid increase in online digital data, it has become very difficult to extract specific information from the internet. In order to overcome this problem, more research efforts are being made on information extraction from unstructured text, e.g. WEB pages and blogs etc. These efforts go beyond any language barriers. The process of finding and extracting information from unstructured data is called Information Extraction (IE). Mainly, IE focuses on identifying *Entities*, *Events*, and *Relationships* between entities [1, 2]. Now, the process of identifying and extracting entities has been well researched and is more stable than the *Events* and *Relationships* extraction [3]. Factors like ambiguity of natural language (NL) and human ability of expressing the same fact with many different ways make IE nontrivial. Formally, Event Extraction (EE) is the process of recognizing events from unstructured text with all involved entities, including any occurring relations [1, 2, and 3]. In other words, EE's final outcome should be able to answer the questions like *Who* (who caused or participated in an event), *When* (when did that event happen) and *Where* (Place of the event) etc. According to the Automatic

[1] PhD Student, Abdel Alnasser Alasfour, Computer Science Department, University POLITEHNICA of Bucharest, Romania, e-mail: Nasser_Asfour@yahoo.com

[2] Prof., Computer Science Department, University POLITEHNICA of Bucharest, Romania, e-mail: stefan.trausan@cs.pub.ro

Content Extraction (ACE) evaluations [2], the event extraction task must contain *Entity Identification*, *Mention*, *Event Trigger*, *Event Arguments* and *Event Mention.* These terminologies are discussed in detail in the later sections. Recently, many events are spread instantly on the internet from all over the world. Most of these events emerge as part of a multi-event hierarchy or as standalone events. During the last four years, a major event (The Arab Spring) has been seen in the Arab world. Since the beginning of this event, a large number of people have been trying to follow and learn any developments happening in relation to that event. In this article, we present a new approach for automatic multilingual event extraction from text (OraMEE). In the subsequent sections, we demonstrate OraMEE creation, starting by setting up the proper template for holding all the extracted events and identifying OraMEE interest domain (the Arab Spring events). Rest of the article is organized as follows:

Section II defines and presents major components of an IE system. Section III briefly talks about the MLIE and CLIR systems. Section IV demonstrates Event Extraction method; whereas, section V describes the OraMEE framework environment. Section VI talks about the proposed approach. Section VII analyses and evaluates the OraMEE approach and section VIII concludes the article.

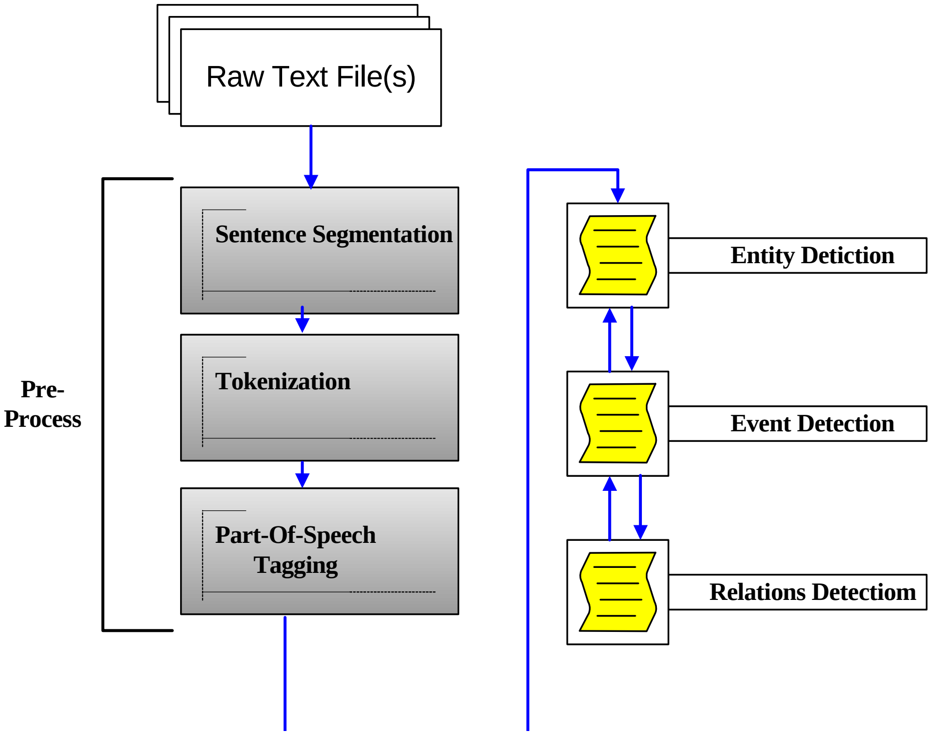


Fig. 1-Inormation Extraction systems main components

2. Information Extraction (IE)

Basically, IE is the process of automatic analysis of the unstructured text in order to identify, locate and sometimes predict classes of Named Entities (NE), Relations between these entities and Events [2, 3].

Any IE system is a domain driven system with a pre-defined template to control the outcomes format. Fig. 1 demonstrates the IE system's major components, which are:

- Pre-Process components including
    1. Sentence segmentation.
    2. Tokenization
    3. Part-Of-Speech Tagging
- Named entity identification
- Events identification
- Relations identification

Traditionally, texts files are superior sources for information and knowledge, and there are three main methods for extracting information from them [3]. These classifications are made on basis of the algorithm used to extract information in general, and event extraction in particular and they are:

1- **Data-driven approaches.**
These approaches essentially rely on quantitative methods to find out relationships within texts by implying statistics, probabilistic approaches to automated language processing such as probabilistic modeling and information theory. All these methods mainly concentrate on discovering statistical relations. In [7, 8, 9, 10], data-driven approaches have been used for event extraction from the unstructured data.

2- **Knowledge-driven methods.**
In general, Knowledge-driven methods rely on rule-based patterns for representing knowledge. For that reason, they essentially employ different linguistic lexicographic information combined with human knowledge to express the information in the processed text. Predefined patterns are useful in the case of extracting very particular information.

3- **Hybrid event extraction approaches.**
A combination of Data-Driven and knowledge-driven approaches can give more acceptable results in terms of information retrieval.

3. Multilingual Information Extraction (MLIE)

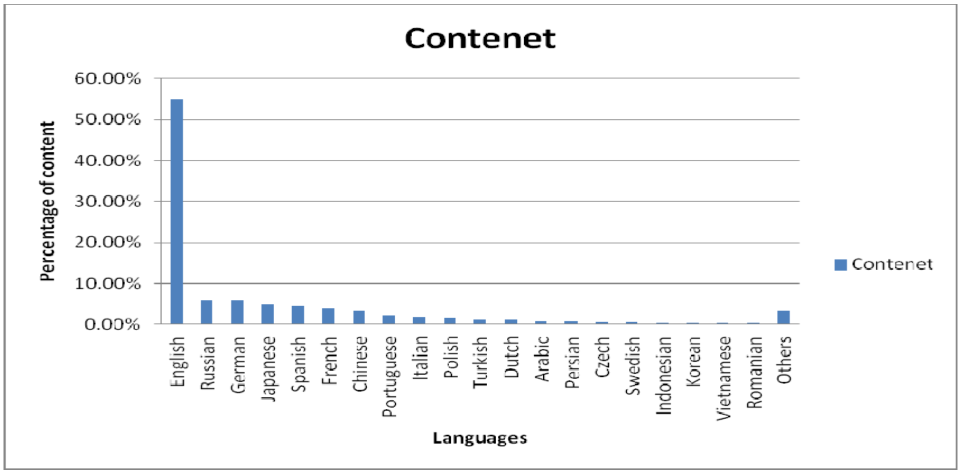


Fig. 2-Represent the internet content percentage per languages- Internet World Stats

The Internet has become a huge international information encyclopedia by supporting many different languages and covering most of the life aspects. Also, it has become a huge container for different digital contents like books, pictures, videos, and so on. Currently, internet has become as a media tool for knowledge and cultural exchange between people from different nations. Nowadays, Internet is considered to be the easiest and fastest way to get to the information. In fact, statistics reveal that the language proportion of digital content is different from actual proportion, e.g. English content rate in the internet is 55.7%, while Arabic language rate is 0.3%. Same holds true for other languages like Russian 6% and Japanese 5% etc. Fig. 2, shows each language’s share in the global digital content [13]. Obviously, not every internet user is able to read or write English, and most of the English content is not available for other languages.

In summary, most of the internet users lose access to huge information, due to language constraints. In order to overcome the language limitations for the digital world, many ideas have been floated, e.g. Cross Languages Information Retrieve (CLIR) and Multilingual Information Extraction (MLIE) etc. CLIR is an NLP task in which the system retrieves and classifies a relevant document, despite its content language. In order to get the job done, CLIR systems follow two different methods, Query Translation and Documents Translation. Fig. 3 demonstrate the query translation method [1, 5].

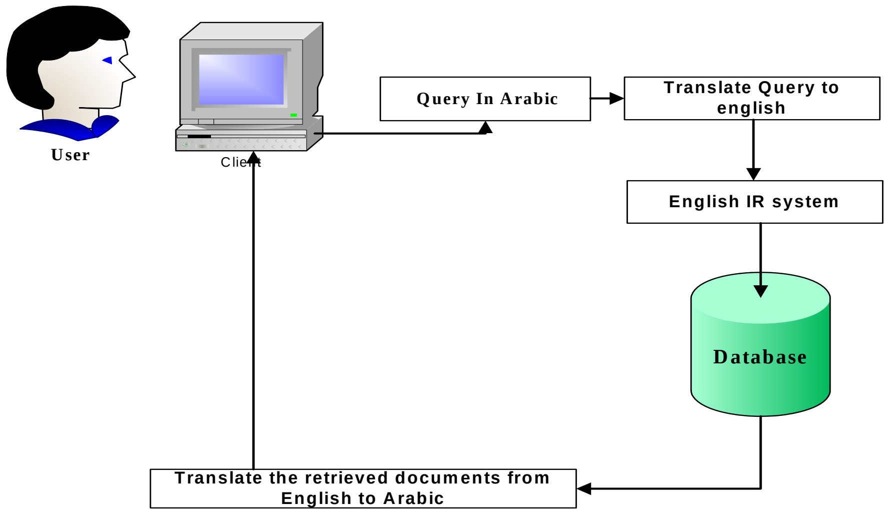

User
Client
Query In Arabic
Translate Query to english
English IR system
Database
Translate the retrieved documents from English to Arabic

Fig. 3- CLIR - Query Translation Method

On the other hand, *Multilingual Information Extraction* (MLIE) is the task of extracting information from non-structured documents, written in different languages to generate structured multilingual representations of their content. MLIE mainly extract information about entities, events and relations [1, 2, 3]. In other words, MLIE systems are the same as any normal IE system, except MLIE systems have an additional feature of extracting information from multilingual documents.

4. Event Extraction

In general, "event" has many definitions and meanings. One of the most acceptable definitions is "*Occurrence happening at a determinable time and place, with or without the participation of human agents*" [6]. In addition, an event may occur in succession of other event occurrences. On the other hand, the definition of event extraction from the IE point of view is to automatically identify events within text and generate proper information about them, to enable identification of each event for the four Ws, which are *Who*, *What*, *When*, and *Where*. Furthermore, some systems can exceed their capabilities to identify more attribute such as *What* (tools) and *Why* (reasons). The task of Event Extraction (EE) has been a point of interest in the IE discipline, since the first declaration in the year of 1987 at the **Message Understanding Conferences (MUCs)** [3]. In the beginning, MUCs focused on limited topics such as terrorist activities and management succession; also they depend on a predefined template for extracting the detected events [2]. Since then, many EE systems have been proposed on MUC's guidelines, e.g. extracting infectious disease events, and natural disaster events [7, 8, and 10]**.** Essentially, locating and detecting event in any sentence needs enormous efforts; because, an event can be expressed within or beyond the sentence boundaries. Furthermore, event information can be distributed over multiple sentences. Therefore, there are two approaches to detect and locate events. First is a *Structural transformation,* in which we can find all the events having the same synonym meaning as the target event. Second is the *Evidence combination* method, in which the system identifies a new event infers from a collection of events [1, 2, and 3]. Mostly, any event extraction system must contains these Process and components in its architecture:

1- Identify Named Entity Reorganization and the relations between them.
2- Identify and detect sentence structures (nouns, verbs etc.) and their boundaries.
3- Detect and resolve synonyms.
4- Combine sentences when the event information is disseminated across multiple sentences.
5- Re-present events with a new inferences event.

Furthermore, the Automatic Content Extraction (ACE) has set the following list of terminology for any event extraction task:

Entity*: an object or a group of objects, such as Persons, Organization, Cities etc.*
Mention*: the word which clarify the entity.*
Event trigger*: The word which point an event.*
Event arguments*: the involved entities in event trigger*
Event mention*: the tagged information for an event.*

On the other hand, ACE has set a list of eight main events categories, and each category contains different events subtypes:

1- *Life* (Be-Born, Marry, Divorce, Injure, Die etc.).
2- *Movement* (Transport).
3- *Transaction* (Transfer-Ownership, Transfer-Money etc.).
4- *Business* (Start-Org, Merge-Org, Declare-Bankruptcy, End- Org etc.).
5- *Conflict* (Attack, Demonstrate etc.).
6- *Contact* (Meet, Phone-Write etc.).
7- *Personnel* (Start-Position, End-Position, Nominate, Elect etc.).
8- *Justice* (Arrest- Jail, Release-Parole, Trial-Hearing, Charge-Indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon etc.).

5. OraMEE Framework Environment

In order to build our application for events extraction, we use three Oracle integrated tools which are Oracle Text, Developer 6i and PL-SQL developer. Oracle11g is used as a database engine to hold and manages OraMEE schema.

A. Oracle Text: Oracle Text package is part of Oracle11g and it is build especially for dealing with text [11]. It contains the proper functions to

power the users in order to index and analyze different types of documents, such as TEXT, PDF and XML files etc.
Oracle in general and Oracle Text packages, functions and procedures in particular provide an ideal environment to perform different linguistic analysis; due to the integrated full text retrieval features, such as:

1- Variety of text searching options.
   - ✓ Keyword searching.
   - ✓ Contextual queries.
   - ✓ Boolean operations.
   - ✓ Pattern matching.
   - ✓ Mixed thematic queries.
   - ✓ HTML/XML section searching.

2- Supports multiple languages.

3- Advanced document relevance

4- Advanced documents ranking technology.

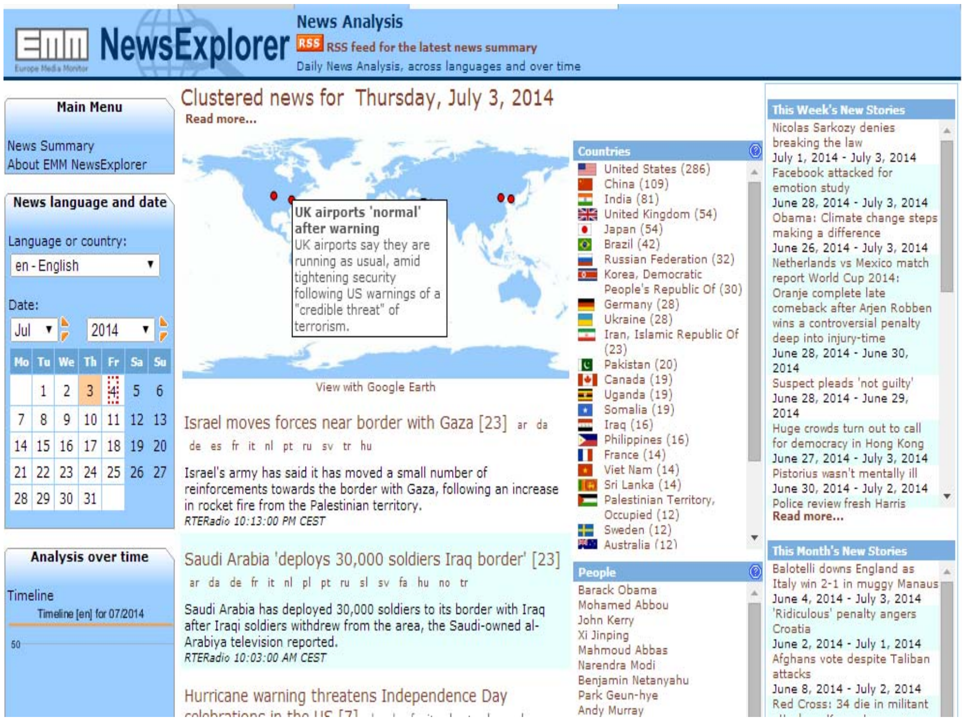
B. Oracle Developer 6i: This tool is used to create the graphical user interface (GUI) for OraMEE application including forms and reports.

C. PLSQL Developer: PLSQL is an oracle standalone tool, and mainly it is used to simplify the development and management of OraMEE schema including tables, indexes, functions and procedures.

6. Related Application

The related application can be divided into several categories depending on the techniques they apply in their approaches.

A. **Clustering Techniques:** The main technique is *clustering* the document or the news articles by the type of the event contained in the document. An example of these applications is NewsExplorer [16]. Main goal for NewsExplorer is to cluster news with the same events into one group. At the same time, it also provides news clustering on several levels such as locations, person’s names, and time period see Fig. 4.



Mo	Tu	We	Th	Fr	Sa	Su
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

Clustered news for Thursday, July 3, 2014

UK airports 'normal' after warning
UK airports say they are running as usual, amid tightening security following US warnings of a "credible threat" of terrorism.

Israel moves forces near border with Gaza [23] ar da de es fr it nl pt ru sv tr hu

Israel's army has said it has moved a small number of reinforcements towards the border with Gaza, following an increase in rocket fire from the Palestinian territory.
RTERadio 10:13:00 PM CEST

Saudi Arabia 'deploys 30,000 soldiers Iraq border' [23] ar da de fr it nl pl pt ru sl sv fa hu no tr

Saudi Arabia has deployed 30,000 soldiers to its border with Iraq after Iraqi soldiers withdrew from the area, the Saudi-owned al-Arabiya television reported.
RTERadio 10:03:00 AM CEST

Hurricane warning threatens Independence Day celebrations in the US [7]

Countries

United States	(286)
China	(109)
India	(81)
United Kingdom	(54)
Japan	(54)
Brazil	(42)
Russian Federation	(32)
Korea, Democratic People's Republic Of	(30)
Germany	(28)
Ukraine	(28)
Iran, Islamic Republic Of	(23)
Pakistan	(20)
Canada	(19)
Uganda	(19)
Somalia	(19)
Iraq	(16)
Philippines	(16)
France	(14)
Viet Nam	(14)
Sri Lanka	(14)
Palestinian Territory, Occupied	(12)
Sweden	(12)
Australia	(12)

People

Barack Obama
Mohamed Abbou
John Kerry
Xi Jinping
Mahmoud Abbas
Narendra Modi
Benjamin Netanyahu
Park Geun-hye
Andy Murray

This Week's New Stories

Nicolas Sarkozy denies breaking the law
July 1, 2014 - July 3, 2014
Facebook attacked for emotion study
June 28, 2014 - July 3, 2014
Obama: Climate change steps making a difference
June 26, 2014 - July 3, 2014
Netherlands vs Mexico match report World Cup 2014: Oranje complete late comeback after Arjen Robben wins a controversial penalty deep into injury-time
June 28, 2014 - June 30, 2014
Suspect pleads 'not guilty'
June 28, 2014 - June 29, 2014
Huge crowds turn out to call for democracy in Hong Kong
June 27, 2014 - July 3, 2014
Pistorius wasn't mentally ill
June 30, 2014 - July 2, 2014
Police review fresh Harris
Read more...

This Month's New Stories

Balotelli downs England as Italy win 2-1 in muggy Manaus
June 4, 2014 - July 3, 2014
'Ridiculous' penalty angers Croatia
June 2, 2014 - July 1, 2014
Afghans vote despite Taliban attacks
June 8, 2014 - July 2, 2014
Red Cross: 34 die in militant

Fig. 4 - Screenshot form NewsExplorer online System, http://emm.newsexplorer.eu

B. **Summarization Techniques:** Another example for event extraction is Global Event-Data System (GEDS). "The GEDS has been established to allow computer-assisted identification, narrative description and analytical coding of daily international and intra-national events, as reported primarily in on-line news sources" [2]. In order to collect an event, GEDS mainly focuses on English news article's headline and then create automatic summary for that article, then the system stores that summary into database table.

C. **Manuals Event Collection Approaches:** In this approach, the event was collected and categorized by human experts. The experts read and search the media, and try to find any related events. Such documents are inserted in a suitable format for further analysis and evaluation. As an example of such systems is the Conflict and Peace Data Bank (COPDAB). According to [2], the COPDAB's main goal is to analyze events that happened within a the time of World War II.

7. Oracle Multilingual Event Extraction Approach (OraMEE)

As depicted in Fig. 5, OraMEE contains many components as. Furthermore, OraMEE follows the ACE slandered definition for event extraction when come to events extractions task. In general, any IE system must have a template to format the output (section II). Furthermore, IE system must be a domain driven system.

1- **OraMEE Template includes**.
   - ✓ **Event:** any words or phrase that declare an event
   - ✓ **Date-Time stamp:** the date and time for that event
   - ✓ **Event Status:** shows the status of that event (Open, closed, not available –NA etc.)
   - ✓ **Place:** the event's geographical information (City, Street, Hall etc.)
   - ✓ **Participants:** the involved entities (Persons, Companies, etc.)
2- **OraMEE Domain**: The selected domain for OraMEE are the events that come from the Arabic Spring, and are extracted from special corpus (EventCorp), which is created by collecting and indexing most of the Arab Spring event spreads over the internet [4]. EventCorp contains two languages (Arabic and English), and it is available for research.
3- **Event Extractor Approach (EEA).**

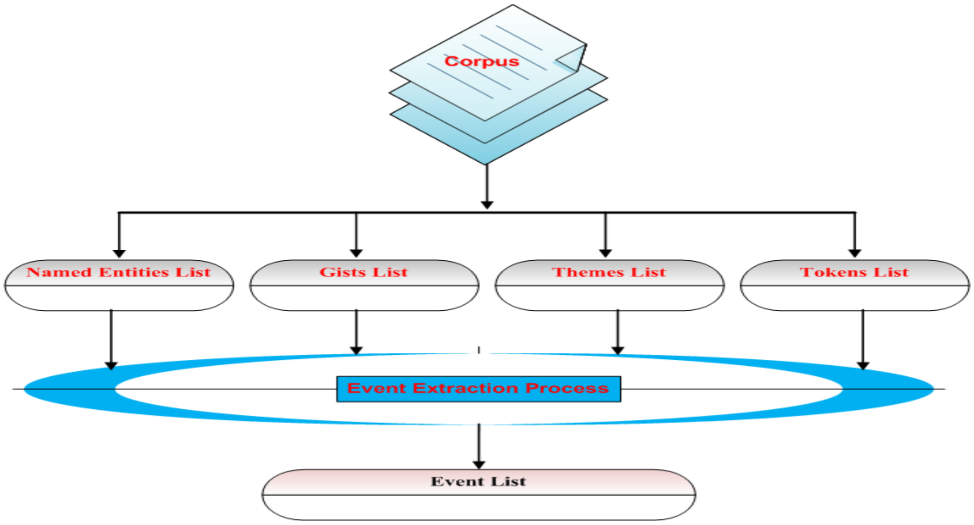
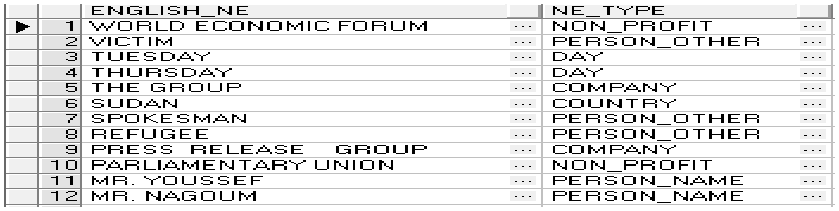


Fig. 4- Event Extraction Approach

OraMEE event extraction mainly depends on four lists:

- ✓ **Named Entities List.**
  In order to extract named entities from the corpus, Oracle Named Entity Recognition and extraction (NER) package is used for NE recognition, extraction, and finally classifying them into different

categories like Company, Person Name, Country etc., Fig. 6 shows a sample NE list.



		ENGLISH_NE	NE_TYPE
▶	1	WORLD ECONOMIC FORUM	NON_PROFIT
	2	VICTIM	PERSON_OTHER
	3	TUESDAY	DAY
	4	THURSDAY	DAY
	5	THE GROUP	COMPANY
	6	SUDAN	COUNTRY
	7	SPOKESMAN	PERSON_OTHER
	8	REFUGEE	PERSON_OTHER
	9	PRESS RELEASE GROUP	COMPANY
	10	PARLIAMENTARY UNION	NON_PROFIT
	11	MR. YOUSSEF	PERSON_NAME
	12	MR. NAGOUM	PERSON_NAME

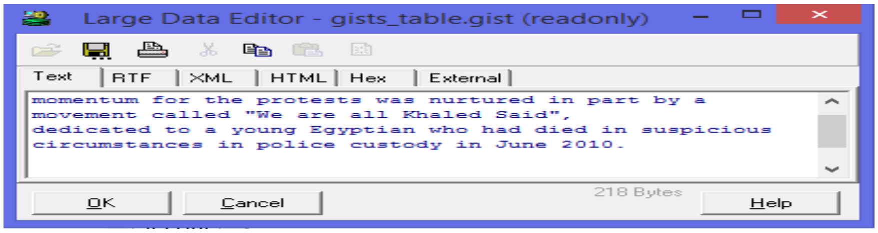
Fig. 5- Named Entities sample list

✓ **Gist List.**

A gist is the document essence or the document summary. In most cases, the Gist is a paragraph or more, which describes the most content of the whole txt. In order to extract and build a gist for each document in the target corpus, an oracle Gist module is used. Fig. 7 shows a Gist sample extracted over a new article. Oracle Gist module is built and integrated with a huge semantic ontology, known as knowledge base. It supports two different languages, which are English and French. In addition, Oracle Gist module has the ability to extract and build two kind of Gist

1. **Generic gist**: a paragraph or more for representing each document in the corpus.

2**. Point of View (POV):** each document in the corpus is tagged with one word or more as a POV.



```
momentum for the protests was nurtured in part by a
movement called "We are all Khaled Said",
dedicated to a young Egyptian who had died in suspicious
circumstances in police custody in June 2010.
```

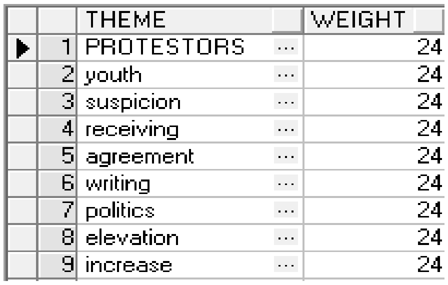
Fig. 6 - A Gist sample extracted from a document

✓ **Themes list.**

A theme is the main impression or main idea, which abstracts the whole document. In other words, theme tries to answer the question of

"***What is this document talking about***?"

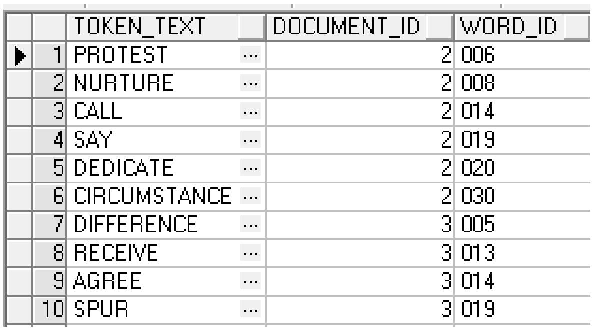
Like Gist module, Oracle also has another build-in module special for creating themes over documents. Fig. 8 represent a list of themes and their weight as a rank for best describing the document.



		THEME	WEIGHT
▶	1	PROTESTORS	24
	2	youth	24
	3	suspicion	24
	4	receiving	24
	5	agreement	24
	6	writing	24
	7	politics	24
	8	elevation	24
	9	increase	24

Fig. 7 - Themes list sample with wieght for each theme

- ✓ **Tokens List :**
  As a result of corpus tokenization, a list of tokens will be produced for all the documents within the corpus, see Fig. 9 for a sample of tokens extracted from a text.

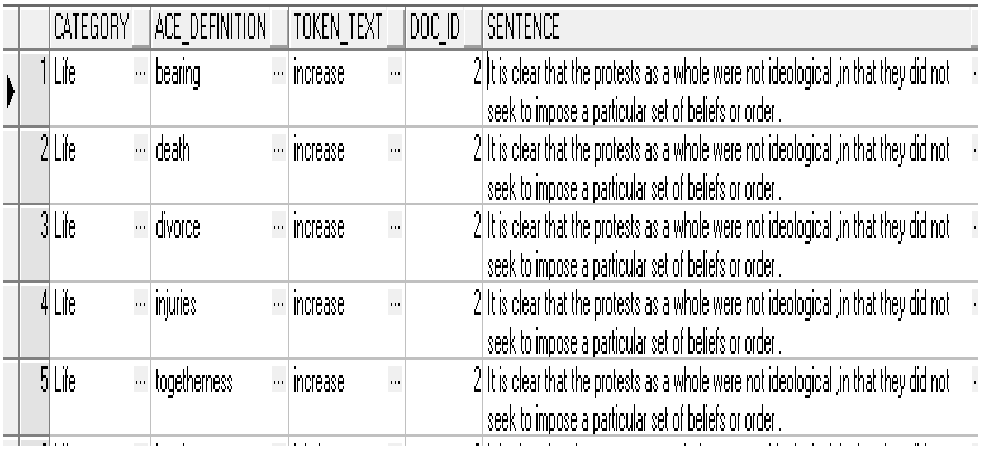


		TOKEN_TEXT	DOCUMENT_ID	WORD_ID
▶	1	PROTEST	2	006
	2	NURTURE	2	008
	3	CALL	2	014
	4	SAY	2	019
	5	DEDICATE	2	020
	6	CIRCUMSTANCE	2	030
	7	DIFFERENCE	3	005
	8	RECEIVE	3	013
	9	AGREE	3	014
	10	SPUR	3	019

Fig. 8 - tokens list sample

- ✓ **Event extraction process.**

  By combining the outcomes from each list, Named Entity list, Gist list, themes list and tokens list, the proposed system builds an initial event list as shown in Fig. 10.



	CATEGORY	ACE_DEFINITION	TOKEN_TEXT	DOC_ID	SENTENCE
1	Life	bearing	increase	2	It is clear that the protests as a whole were not ideological ,in that they did not seek to impose a particular set of beliefs or order .
2	Life	death	increase	2	It is clear that the protests as a whole were not ideological ,in that they did not seek to impose a particular set of beliefs or order .
3	Life	divorce	increase	2	It is clear that the protests as a whole were not ideological ,in that they did not seek to impose a particular set of beliefs or order .
4	Life	injuries	increase	2	It is clear that the protests as a whole were not ideological ,in that they did not seek to impose a particular set of beliefs or order .
5	Life	togetherness	increase	2	It is clear that the protests as a whole were not ideological ,in that they did not seek to impose a particular set of beliefs or order .

Fig. 9 - Event initial list

4- OraMEE Approach

In order to create an automatic event extraction approach, we depend on a hybrid IE method and a combination of IR and IE technology. Fig. 11 demonstrate OraMEE approach:

1. Accept user query (Arabic)
2. Translate user query (to English, French)
3. Search the corpus.
4. If no relevant document found, then go to 1, else continue
5. Sentences segmentation
6. Indexing (tokenization)
7. Entities extraction
8. Translate entities to Arabic
9. Event extraction (**EEA**)
10. Translate events to Arabic
11. Fill-in Arabic, English and French event template table.
12. End.

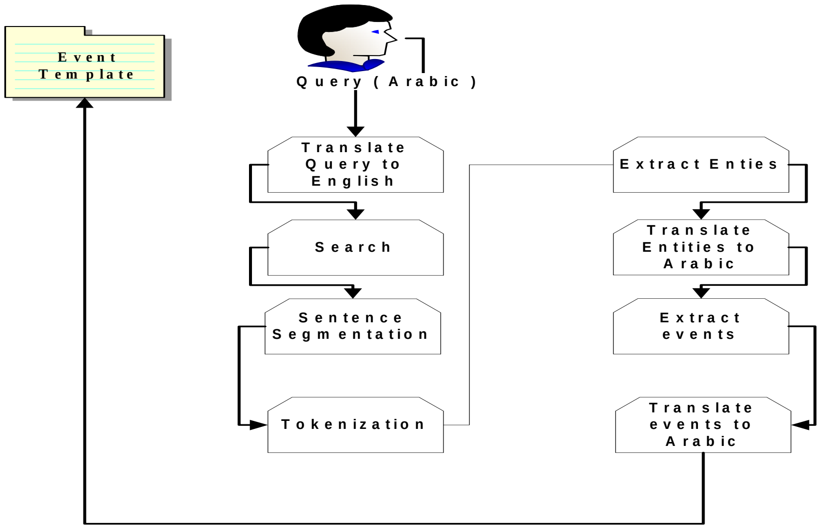


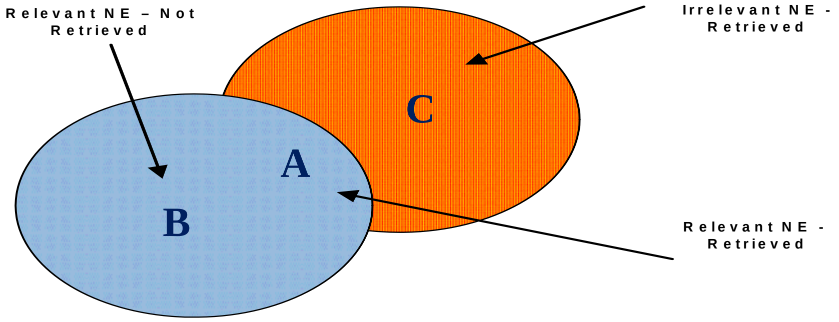
Fig. 10- OraMEE main components

5-. OraMEE main components

a. The translation is done automatically on the runtime by integrating Google translate API with OraMEE core [12]
b. Entities recognition and extraction is done by Oracle Named entities function.
c. Event recognition and extraction is done by special procedure, built by mixing the functionality of Context index NEAR function, and it depends on ACE standards for event extraction.
d. The English version from EventCorp is uploaded to oracle database as a training (learning OraMEE) corpus

8. OraMEE Evaluation

In order to evaluate OraMEE Approach, we perform the best known statistical measurements in information retrieval task, namely **Recall** and Precision. Recall is the ratio of retrieved relevant *EVENT* to the total number of existing relevant *EVENTs*; whereas, **Precision** is the ratio of retrieved relevant *EVENTs* to the whole retrieved *EVENTS* [14]. These measurements are well known for validation of any information retrieval activities; especially, to test the retrieved information relevance's. Fig. 12 presents Recall, Precision, and their formula.



$$\text{Recall} : \frac{A}{A + B} \qquad \text{Precision} : \frac{A}{A + C}$$

Fig. 11- Recall and precision demonstration and their formulas

Furthermore, an annotated corpus is created for the purpose of evaluating OraMEE approach. The corpus contains a list of 3563 text, all of them in the conflict domain. These texts are collected from the web, mainly from ACLED (Armed Conflict Location and Event Data Project) site [15]. Then these text files are filtered and annotated (manually) in order to get high accuracy results. The final corpus file list characteristics are represented in table 1.

Table 1

The characteristics of Annotated corpus.

Number of files	Words types	Word tokens	Tokens/files Ratio
3563	5445	58532	17

The annotated corpus contains many event instances like "Attack" and "Clash", "Demonstrate" and "Protest" and also pure conflict event. Table 2, shows the statistics result for each the combined event pair and the conflict event.

For the purpose of evaluation OraMEE, another copy of corpus texts is saved without any annotation marks. As mentioned earlier, OraMEE has two approaches, direct and thematic approach.

Table 2

The Statistics result of the annotation corpus

Event sub type	count
Attack&clash	1099
Demonstrate&Protest	2092
Conflict (general)	372

In the direct approach, OraMEE is able to identify 1847 from all the included events, in addition to 221 indirectly declared events like "Assailants" and "siege". The final result is depicted in table 3.

Table 3

The final results for evaluating OraMEE

TP	FP	FN	Precision	Recall	F-score
2068	0	1495	0.58	1	0.73

11. Conclusion and Future Works

To the best of our knowledge, no other research work has combined Themes, Gist and ACE list to extract and create an event representation template. Simple Implementation of a new approach for event extraction on basis of the document Gist and Themes, combined with Named entities and ACE categories of the event is a promising method. Such a technique can be used for other languages as well. In future work, this approach will be used as a core component for multilingual event extraction system, built over highly technical integration tools like Oracle Text. OraMEE in the direct methods was able to identify a set of 1847 from the total of 3563 events instances. The reason behind this is, OraMEE has a problem in identifying the different shapes of the same verb; for example, the event Attack can appear in different situations such as "Attacked" or "Attackers" or "Attacking" or just attack and so on. In future work, we will add a new component to stem the tokens in the tokens list before running OraMEE approach. On the other hand, OraMEE was able to identify undeclared events such as "Assailants" and "siege" by building themes and gist for the corpus text. We hope when adding the mentioned component(stems), OraMEE will be able to identify and extract more hidden events and be able to build an acceptable event representation template.

REFERENCES

[1]. TIDES, 2004, DARPA Program in Translingual Information Detection Extraction and Summarization [http://tides.nist.gov/].

[2]. LDC, 2004, ACE: Automatic Content Extraction [https://www.ldc.upenn.edu/].

[3]. NIST, 1999, MUC: Message Understanding Conference [http://www.itl.nist.gov/]

[4]. Alasfour, A., Trausan-Matu, S., 2013 , Developing an Arabic Corpus for Event Mining , ICSTCC2013, p. 21-28 ,Oct 11-13, 2003, Sinaia, Romania.

[5]. Huttunen, S., Yangarber, R., Grishman, R.: Complexity of event structure in information extraction Scenarios. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002). Taipei (2002)

[6]. Business Dictionary [Online], accessed on 01-05-2014, available at : http://www.businessdictionary.com/ .

[7]. Ralph Grishman, Silja Huttunen, and Roman Yangarber. Real-time event extraction for infectious disease outbreaks. In Proceedings of the second international conference on Human Language Technology Research, pages 366–369, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[8]. Martin Atkinson, Jakub Piskorski, Bruno Pouliquen, Ralf Steinberger, Hristo Tanev, and Vanni Zavarella. Online-monitoring of security-related events. In COLING '08: 22nd International Conference on on Computational Linguistics: Demonstration Papers, pages 145–148, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

[9]. Okamoto, M., Kikuchi, M.: Discovering Volatile Events in Your Neighborhood: Local-Area Topic Extraction from Blog Entries. In: 5th Asia Information Retrieval Symposium (AIRS 2009). Lecture Notes in Computer Science, vol. 5839, pp. 181{192. Springer-Verlag Berlin Heidelberg (2009).

[10]. Kim and Rebholz-Schuhmann Journal of Biomedical Semantics 2011, 2(Suppl 5):S3 http://www.jbiomedsem.com/.

[11]. Cathy Shea, Oracle Text Reference, 11g Release 2 (11.2).Part No. E24436-04. Oracle Corporation. Feb. 2014: http://docs.oracle.com/

[12]. Google Translate API [Online].Available at : google.com/translate/

[13]. Internet World Stats [Online]. Available at: http://www.internetworldstats.com/

[14]. William E. Underwood, Matthew G. Underwood , Evaluation of Document Retrieval Technologies to Support Access to Presidential Electronic Records PERPOS Technical Report ITTL/CISTD 02-3,December 2002. http://perpos.gtri.gatech.edu/

[15]. Raleigh Clionadh, Andrew Linke, Håvard Hegre and Joakim Karlsen.(2010). Introducing ACLED-Armed Conflict Location and Event Data. Journal of Peace Research 47(5) 1-10.

[16]. Steinberger Ralf, Bruno Pouliquen, Camelia Ignat (2005). *Navigating multilingual news collections using automatically extracted information*. Journal of Computing and Information Technology - CIT: 13.4, pp. 257-264.

[17]. Dvies John L., (1998), *The global event-data system 1998 Revision [online]*. Center for International Development and Conflict Management and Department of Government and Politics, College Park MD, Available at: [http://www-rohan.sdsu.edu/GEDSCodebook800_3.pdf].

[18]. Azar Edward, (1980). *The Conflict and Peace Data Bank (COPDAB) Project*. Journal of Conflict Resolution 24: 143-252.