KEYS: A Knowledge Extraction System Based on UNL Knowledge Infrastructure

Sameh Alansary[*1], Magdy Nagi[**2]

[*]*Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria,*
[1]sameh.alansary@bibalex.org

*** Computer and System Engineering Department, Faculty of Engineering, Alexandria University, Alexandria, Egypt*
[2]magdy.nagi@bibalex.org

Abstract: ***With the revolution of information available on the internet pages, humans need to extract specific information. This paper presents KEYS (Knowledge Extraction sYStem); an information retrieval and extraction system. It searches for information inside documents represented in UNL, i.e., in semantic hyper-graphs. This allows for retrieval and extraction practices that are language-independent and semantically-oriented. It is expected to provide high-quality knowledge extraction through a shallow analysis of the source text into the Universal Networking Language (UNL) using a specific ontological relations and fully-automatic generation from the resulting UNL document into several different target languages. This is expected to present a novel approach to the topic of identifying the named entity; extracting names with all its types from a natural language form.***

1 INTRODUCTION

Information extraction (IE) is the process of scanning text for information relevant to some interest, including extracting entities, relations, and, most challenging, events{or who did what to whom when and where}. It requires deeper analysis than keyword searches, but its aims fall short of the very hard and long-term problem of text understanding. Information extraction technology arose in response to the need for efficient processing of texts in specialized domains. For example, an information extraction system designed for a terrorism domain might extract the names of perpetrators, victims, physical targets, weapons, dates, and locations of terrorist events. An information extraction system designed for a business domain might extract the names of companies, products, facilities, and financial figures associated with business activities. Full-sentence parsers expended a lot of effort in trying to arrive at parses of long sentences that were not relevant to the domain, or which contained much irrelevant material, thereby increasing the chances for error. Information extraction technology, by contrast, focuses in on only the relevant parts of the text and ignores the rest [1].
Message Understanding Conferences (MUC) have described IE as consisting of different tasks. These various tasks differ mainly in their complexity degree and in the depth of the extracted information. For instance, the named entity (NE) task identifying within free text, person, location and organization names, and quantities, such as dates, monetary amounts, etc. Then a more complicated task which is the coreference task (CO) that involves the identification of coreferent entities in text. The template elements (TE) task is responsible for discovering specific attributes about these entities. Next, the relation extraction (RE) task which implies the detection of specific relations (such as employee of, author of, etc.) within the identified entities. Finally, the most complex task which is the scenario template (ST) task in which the system is required to identify instances of a specific predefined event in the text, and extract the information related to each instance of the found event. The system is expected to provide an event template containing various pieces of event information corresponding to each event detected within the given text. Thus, locating the various forms of interesting information embedded in free text is highly complicated [2].
knowledge extraction has originated from people`s need to obtain and manage the vast amounts of information described in free text more accessibly. Free text contains a multitude of information such as (name of people, places, organizations, roles played by entities in events, relations between entities, etc) that if effectively extracted, can be of great use to many real-world text/web applications, for example, integration of product information from various websites, question answering, contact in formation search, finding the proteins mentioned in a biomedical journal article, and removal of the noisy data [2].

Reference [3] has summarized some of the early work done in the field of information extraction. They have mentioned the work or reference [4], [5] who has analyzed news stories as one of the early attempts in the field of IE. The system is called FRUMP; it is a general purpose NLP system designed to analyze news stories and to generate summaries for users logged into the system. This system is very similar to the current IE systems, since the generated summaries are essentially event templates filled in by FRUMP and presented as single sentence summaries of the events. FRUMP uses hand-coded rules for 17 “prediction” and “substantiation” (the two components of the system) to identify role fillers of 48 different types of events. FRUMP uses a data structure called “sketchy script”, which is a variation of “scripts” that was previously used to represent events or real-world situations described in text [6], [7]. This era of the field of IE has included other approaches as the Prolog-

based system by Silva and Dwiggins [8] for identifying information about satellite-flights from multiple text reports. Cowie [9] has also implemented a system, based on Prolog that uses “sketchy syntax” rules to extract information about plants. By segmenting the text into smaller parts, depending on pivotal points, like pronouns, conjunctions, punctuation marks, etc., the system can avoid the need for complex grammars to parse texts. Sager [10] has also developed a system which is applied to highly domain-specific medical diagnostic texts (patient discharge summaries) to extract information into a database for later processing. The system uses English grammar rules to map the text into a structured layout. Zarri`s work is also worth of noting whose goal was to identify information about relationships and meetings of French historical personalities and represent this information in a structured form in the “RESEDA semantic metalanguage” [11]. The system uses rules for semantic parsing and heuristic rules of identifying slot-fillers required by the RESEDA metalanguage. During this period of NLP research, IE has been a field of interest where a fair amount of efforts have been exerted. Much of this work focused on specific domains, used hand-crafted rules and did not have standard data sets or standard evaluation procedures. Defense Advanced Research Projects Agency (DARPA) has organized a series of Message Understanding Conferences (MUC)[12] as a competitive task with standard data and evaluation procedures in the late 1980s and early 1990s. DARPA has also introduced another program towards the end of the MUC era, it was called the TIPSTER program [13], [14], [15]. It was designed to advance the state of the art in text processing. Naturally speaking research in IE has continued to grow over the years since MUC and TIPSTER. Moreover, the definition of IE has also gradually broadened to include many different types of information and tasks that differ in their complexity [14]. Many systems, for example, GE [16], SRI [17], UMass [18], NYU [19], etc. have participated in MUC tasks, which considerably helped in the advancement of IE research.

However, the field of Information Extraction (IE) still includes vast potentials for large-scale knowledge acquisition, since the current systems are still unable to form a coherent theory from a textual corpus which involves representation and learning abilities, although, the current IE systems are able to uncover assertions about individual entities with an increasing level of sophistication and text understanding. Compared to individual relational assertions provided by IE systems, a theory includes coherent knowledge of abstract concepts and the relationships among them. Previous efforts in text-based knowledge acquisition can largely be attributed to the field of Information Extraction (IE), where the task is to recognize entities and relations mentioned within text corpora. Traditional IE systems focused on identifying instances of narrow, pre-specified relations, such as the time and place of events, from small homogeneous corpora. Furthermore, the current IE systems are typically designed for a single domain, there is a lot of interest in building systems that are easily applicable to new domains [20].

The KnowItAll system is considered as an advancement in the field of IE by capturing knowledge in a manner that scaled to the size and diversity of relationships that are present within millions of Web pages. It is a system that aims to automate the tedious process of extracting large collections of facts from the web in an autonomous, domain-independent, and scalable fashion. By learning to label its own training examples using only a set of domain-independent extraction patterns and a bootstrapping procedure, KnowItAll has managed to accomplish this task. KnowItAll is capable of self-supervising its training process; however, the extraction is not fully automatic. KnowItAll requires a user to determine the relation before each extraction cycle for every relation of interest. When acquiring knowledge from corpora as large and varied as the Web, the task of anticipating all relations of interest becomes extremely complicated [20].

Some of the previous information extraction tools can deal with Arabic, such as Rocket AeroText and NetOwlExtracto. Both of them are capable of discovering entities (people, products, dates, places, and more) and the relationships between them, as well as sentiment analysis in multiple languages. However, both systems are not free, they were developed as commercial products.

Huge amounts of information in natural language forms exist only in lists of documents and to search all these documents to find just a certain piece of information will be a waste of time. Implementing the Information Extraction techniques in a certain system will with no doubt save a lot of time and efforts while providing precise results. Information Extraction techniques can be used to search various types of documents like historical articles, medical researches and newspapers reports.

Since 1950’s, many research groups have recognized the vital role that the IE plays and started to create projects for tasks like the transformation of a whole encyclopedia to structured forms.

Although these projects have faced some natural language processing problems, modest extraction systems have appeared and have been used in extracting information from a relatively small number of forms. IE technology still needs mature systems in order to match the human performance.

The IE systems usually support one of two approaches either knowledge engineer approach or automatic training approach. In the knowledge engineer approach, after analyzing huge number of natural language data, the designer identifies sets of common patterns for which he develops rules manually that get interpreted by the components of the IE system. However, using this kind

of approach in building the system is considered to be highly time and effort consuming.
In the automatic training approach, there is no need to develop the rules manually, since it depends on implementing a machine learning algorithm in the system which is able to detect and create these rules.
The algorithm must get access to a large number of training texts, these texts have to be annotated manually in order to give the algorithm a sufficient amount of examples which it can learn from and provide the extraction rules [21],[22], [23].

This paper adopts the Universal Networking Language (UNL) framework in building a knowledge extraction system. The aim of UNL is to provide a large collection of semantically annotated texts belonging to different languages. We will present a Knowledge Extraction sYStem named KEYS. It searches for information inside documents that are represented in natural language or UNL expression, i.e., in semantic hyper-graphs. It allows for retrieval and extraction practices that are language-independent and semantically-oriented. With KEYS, we try to start a new fashion in IE by targeting the users aspirations from such application. It is based on a philosophy that is different from the mainstream in the field of IE, since it aims to serve the public which is similar to Google`s goal. Moreover, KEYS originality stems from the fact that it can identify and understand the object depending on its context; it is also able to provide all the suggestions related to this object. KEYS includes SEAN and EUGENE. The former is a shallow enhanced natural language analysis system, it represents natural language texts as semantic networks in the UNL format. While the latter is a natural language generation system, it generates natural language sentences out of semantic networks represented in the UNL format. KEYS is expected to synthesize and normalize the information available on the Web, and to provide summaries extracted out of several different input documents. KEYS has been developed by the Library of Alexandria.

In what follows, section 2 will present the different techniques of information extraction systems. Section 3 sheds light on the project`s history and current status. Section 4 illustrates the basic components of KEYS; the system's open-source components. First, the language resources (dictionaries and grammars). Second, the software used in building and operating the system (analysis and generation engines). Each of these components is described and their current state is specified. Section 5 will describe KEYS`s interface and illustrate how this system is used. In section 6 KEYS`s output will be evaluated. Finally, section 7 will conclude the paper.

2 THE BASIC TECHNIQUES OF INFORMATION EXTRACTION

The basic techniques are pattern matching, lexical analysis, name recognition, syntactic structure, scenario pattern matching, coreference analysis and event merging. These techniques are divided into two main parts. First, all the individual facts are extracted from the documents, these individual facts are integrated together to form larger facts and translated into the required output format, this stage is called the integration phase. Second, coreference analysis is done and inferences are drawn from the explicitly stated facts in the document. The final output of the information extraction is called a template [24]. The first step consists of developing a set of patterns that matches the various linguistic realizations of the individual facts and these patterns are not just sequences of words, they are more complex than that. To develop such patterns many linguistic processes are required starting from lexical analysis and ending with name recognition. Most of the current systems use partial syntactic analysis just to identify the verbal or nominal constituents in the text. After using these general patterns, task specific patterns are used to identify the facts of interest, which are called scenarios according to the Message Understanding Conference (MUC). The second step includes conference analysis and drawing inferences from the explicitly stated facts in the document. At the end the final output from the information extraction is called template.

The Pattern matching is done through matching the text against a set of regular expressions, when a segment of a text (constituent) is matched with one of these regular expressions, the text segment become a label with one or more assigned features. When there is any semantic feature associated with the constituent, they are called events or entities.
In lexical Analysis phase, first, the text is split into sentences then into tokens. Each token is looked up in the dictionary to assign its features and part of speech.
In the name recognition phase, the different types of names and other special forms like currency and amounts are identified and classified. This simplifies the further processing.
Some systems do not have a separate phase for syntactic analysis, others attempt to build a complete parser of sentences. However, most of the systems fall in between by building a shallow parser. Identifying some of the syntactic structure simplifies the extraction of the information or the knowledge. The argument to be extracted often correspond to noun phrase [24]. After dividing the text into syntactic constituents, each constituent has to be associated with some features like the tense, voice and root of the verb in the verbal constituents and for nominal constituents information are associated to the head of the constituent like its number whether it is a proper name or not and so on.
Then, larger nominal phrases are built up by attaching their modifiers to them and in this case these patterns will have some semantic constraints.

In the scenario pattern matching phase, the main target is to extract the main events of the scenario. Then the coreference analysis phase comes next which includes the task of resolving the anaphoric references by searching for the most recent previously mentioned entry of the same type, for example, person if the anaphora was one of the personal pronouns.

In the event merging phase, all the information about an event is collected which may constitute a hard task, because the information may be spread over many sentences. Another problem that may face the extraction systems when collecting information about a certain event is that its information may be implicit and needs to be more explicit.

3 PROJECT HISTORY AND CURRENT STATUS

KEYS is a rule based Knowledge Extraction system; this system requires different linguistic resources and tools with certain features in its background in order to work efficiently. It requires a dictionary that is enhanced with certain features that encompass all the levels of linguistic information whether it is morphological, semantic or syntactic (will be described in details in section 4. It also requires a grammar that is capable of providing an adequate semantic and syntactic analysis. Moreover, it requires tools that exploit these resources. These tools are called SEAN and EUGNE which have been developed in Bibliotheca Alexandrina (will be described in details in section 4. The linguistic resources were developed using the universal networking language within the UNL framework.

The UNL project has been originally proposed in 1996. The responsible organization is the Universal Networking Digital Language (UNDL) Foundation[1]in Geneva, Switzerland [25], [26], [27], [28] and [29]. UNL is the interlingua employed here; it is capable of representing the meaning of the content of natural language texts in an abstract universal format that is not influenced by any language. UNL aims ultimately to allow people to generate, have access to, information and knowledge, in their own native language by breaking down the language barriers that exclude the majority of people from gaining access to information in their native language. The UNL also assumes that any information conveyed by natural language can be formally and usefully represented by semantic networks (sometimes called UNL expression) . In UNL approach; the semantic network must be independent of any natural language in particular (i.e., it must be "universal"). This semantic network is made of three different types of discrete semantic entities: concepts, relations and attributes. Concepts are nodes in the network; relations are arcs linking nodes; and attributes are used to represent information conveyed by natural language grammatical categories (such as tense, mood, aspect, number, etc.) [25] which are a standard set of universally-accessible semantic entities. The semantic network is derived by passing through different stages; tokenization and disambiguation, morphological analysis, syntactic analysis and semantic analysis.

In the UNL framework, the different linguistic levels of analysis are achieved via three types of grammar: N-Grammar, or Normalization Grammar which is a set of rules used to segment the natural language text into sentences and to prepare the input for processing, T-Grammar, or Transformation Grammar which is a set of rules used to transform natural language into UNL or UNL into natural language.

The transformation should be carried out progressively, i.e., through a transitional data structure: the tree, which could be used as an interface between lists and networks. Accordingly, the UNL grammar states seven different types of rules which is divided into two types of grammar; analysis and generation. Three types of rules are common between the two grammars the other four depend on the type of grammar. The seven types of rules are (LL, TT, NN, LT, TL, TN, NT), specified as indicated below:

ANALYSIS (NL-UNL)

- LL - List Processing (list-to-list)
- LT - Surface-Structure Formation (list-to-tree)
- TT - Syntactic Processing (tree-to-tree)
- TN - Deep-Structure Formation (tree-to-network)
- NN - Semantic Processing (network-to-network)

[1]The official website of the foundations is available at http://www.undl.org

GENERATION (UNL-NL)

- NN - Semantic Processing (network-to-network)
- NT - Deep-Structure Formation (network-to-tree)
- TT - Syntactic Processing (tree-to-tree)
- TL - Surface-Structure Formation (tree-to-list)
- LL - List Processing (list-to-list)

Finally, D-Grammar, or Disambiguation Grammar which is a set of rules used to improve the performance of the transformation rules by constraining or forcing their applicability. Grammars are not bidirectional, although they share the same syntax. In the UNLization, the N-Grammar contains the normalization rules for natural analysis, the analysis T-Grammar contains the transformation rules used for natural language analysis and the analysis D-Grammar contains the disambiguation rules used for tokenization as well as for improving the results of the NL-UNL T-Grammar. While in the NLization process, the generation T-Grammar contains the transformation rules used for natural language generation and the generation D-Grammar contains the disambiguation rules used for improving the results of the UNL-NL T-Grammar.

KEYs takes advantage of the UNL approach along with the new trend in NLP applications, that is being an open-source application, because of its vast advantages, opportunities and potentials. A rule-based knowledge extraction system is open source only when the source code of its engines and tools are distributed along with the linguistic data of the extraction pairs. In addition, tools to maintain and develop the linguistic resources so that they can be used with the engines should also be distributed. KEYs fulfills all of the criteria and, hence, can be positively considered an open-source Knowledge extraction system. Moreover, not only its components are open-source, they are also free. The basic components of KEYs, its linguistic resources and tools will be described in details in section 4.

4 THE BASIC COMPONENTS OF KNOWLEDGE EXTRACTION SYSTEM (KEYS)

As mentioned before a knowledge extraction system depends on different linguistic resources and tools in order to be able to operate. KEYS depends on three linguistic resources which are dictionary, corpus and grammar. It also depends on two tools called SEAN and EUGENE. All these resources and tools are developed by Bibliotheca Alexandrina. In this section these resources and tools are going to be described in details.

A. Language Resources

1) *Dictionary:* it presents the linguistic information that constitutes the linguistic infrastructure of the dictionary (the UNL dictionary) used by KEYS application. The linguistic information that appears in the UNL dictionary has been assigned to all of the words of the dictionary through UNLarium[2], encompassing the different linguistic levels: morphological information, morpho-syntactic information, syntactic information and semantic information. UNL uses a standard and universal list of features (Tagset) to describe all types of the linguistic information concerning every natural language word. The words are described using a list of features extracted from the UNDL Foundation Tagset. The UNDL Foundation recommends adopting the following tags for some specific and pervasive grammatical phenomena to boost the standardization of the lexical resources used in the UNL framework. The Tagset's features depending on the structure of the natural language. Several of those linguistic constants have been already proposed in the Data Category Registry (ISO 12620)[3], see Fig. 1

[2] http://www.unlweb.net/unlarium/
[3] http://media.dwds.de/clarin/userguide/text/concepts_ISOcat.xhtml

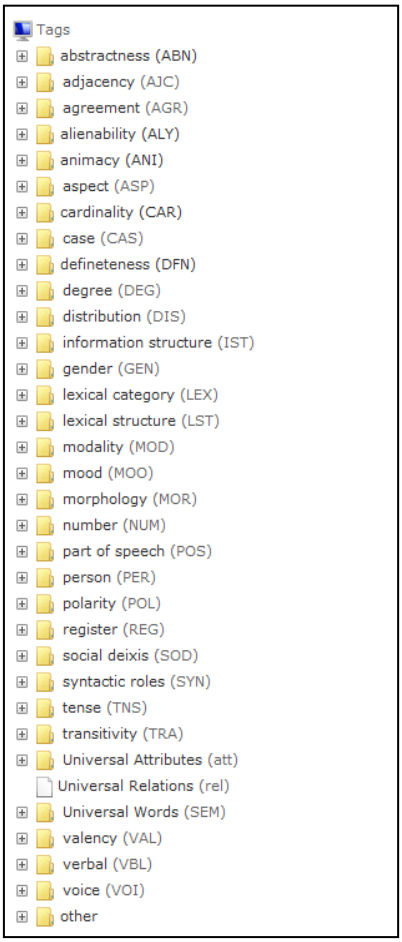

Figure 1: List of tags in alphabetical order

The tagset is providing the technical means for describing any linguistic behavior which should be done in a highly standardized manner, so that others could easily understand and exploit the data for their own benefit. The main intention is to create a harmonized system in order to make language resources as easily understandable and exchangeable. The dictionary is enhanced by morphological information indicating the structure of words, some of this morphological information such as part of speech, lexical structure and the inflections of words.

Part of speech feature: It is used to classify words into main classes and each class may include subclasses. The classes are nouns, verbs, adjective, adposition, adverb, affix, classifier, conjunction, determiner, interjection, numeral, particle and pronoun. The system is designed as such in order to create much flexibility in describing the different types of words. Moreover, the classes are divided into subclasses. For example, the used features in the dictionary differentiate between two types of nouns, common noun such as "صندوق" 'box' - "باب" 'door' – "ورقة" 'paper' and proper noun as "نجيب محفوظ" 'Naguib Mahfouz' - "مصر" 'Egypt' - "اليونسكو" 'UNESCO'.

The dictionary used in the knowledge extraction system differentiates between common and proper nouns and is enhanced with information for the proper names such as the names of rivers, mountains, the names of humans which are considered as public figures (common Arab and non- Arab first and second names).

lexical structure: It is used to classify the words into simple words as the Arabic words "قرأ" 'read' - "مكتب" 'office' - "رائع" 'wonderful', and multiword expressions such as the word "سور الصين العظيم" 'the great wall of China'.

Inflectional paradigms: It is a stored feature that is responsible for generating the different word forms out of the stored lexemes. The dictionary also includes syntactic information that describes the principles and processes by which sentences are constructed. It deals with phrase and sentence formation out of words, such as valency, aspect and sub categorization information. Moreover, the dictionary also is enhanced by information that is concerned with the grammatical categories such as gender, number, person, transitivity, tense, case, voice and mood.

The most important feature concerning building any knowledge extraction system is the semantic classification of the words; the UNL dictionary utilizes a semantic ontology. This ontology classifies the entities existing in the natural world into a semantic hierarchy. This hierarchy points out the particular type of each concept and the kind of relation it indicates with other

concepts in the ontology. Each entry in this hierarchy carries a set of features and attributes and all subclasses of this concept inherit the properties of that class. Ontologies are useful in NLP as they play a crucial role in the disambiguation of word senses as well as the understanding of a natural language text by determining the exact sense of a word via its position in the semantic hierarchy. The semantic ontology adopted in the UNL dictionary is the English WordNet 3.0. ontology. In WordNet, English nouns, verbs, adjectives and adverbs are organized into sets of synonymous words (called synsets), each synset representing one distinct concept. For example, the words "coast", "seacoast", "sea-coast" and "seashore" are all synonyms grouped together in a single synset that refers to a unique cognitive concept which is "the shore of a sea or ocean".
Nouns in the WordNet hierarchy are divided into several semantic fields each having a "unique beginner" as the starting node. A unique beginner is a semantic entity that probably has no hypernym and from which nouns that belong to this distinct semantic field can be pulled out. The WordNet employs a set of 25 unique beginners, 8 of which refer to tangible things or "entities", 5 denote "abstractions" and 3 are "psychological features". Verbs, modifiers and adverbs are also classified into distinct semantic hierarchies see Fig. 2. For more details about the dictionary and the stored features see [30]. Moreover, it is important to mention that any lexical item that is not included in the dictionary will be labeled as "TEMP".

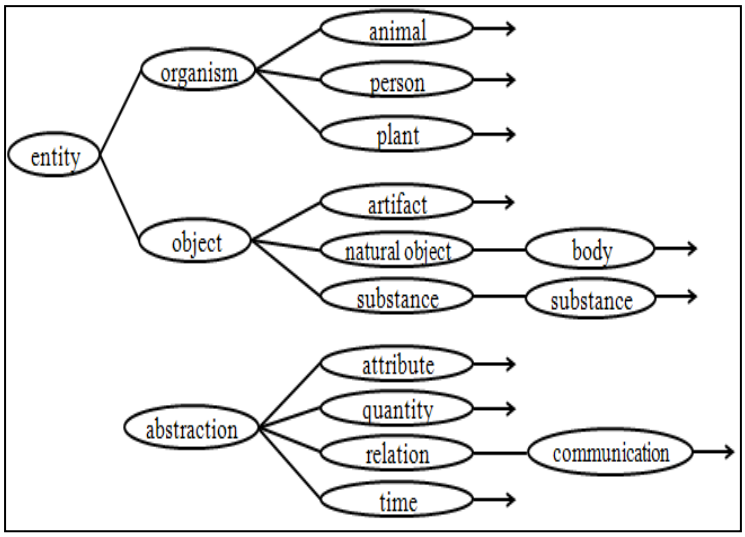

Figure 2: The semantic ontology used in the dictionary

2) CORPUS: In order to build an sufficient corpus for proper names, 1000 pages of proper names have been selected from the Wikipedia. These pages represent a rich material for the corpus, since these pages will include these proper names in real contexts. Furthermore, these proper names are from Wikipedia which means that the coverage rate will be high and the corpus will be considered robust. These pages are segmented into sentences using concordance. The total number of occurrences for these proper names is 22,000 with maximum 7 words length; 3 words before and after the proper name. The data is divided into training data which includes 17,000 occurrences and testing data which includes 5,000 occurrences. The testing data will be used later in the evaluation phase. Fig. 3represents an example for the corpus of the searched word "هدسون" with its 85 instances.

Figure 3: Example for corpus

3) *GRAMMAR:* In KEYS application, the knowledge extraction process has to pass through different stages. Firstly, lexical analysis stage which split sentences into tokens. Each token is looked up in the dictionary to assign its features; part of speech and the other features that are stored in the dictionary. Then a pattern matching stage starts to build ontological relations between the tokens, but if the grammar fails to match the sentences with any pattern, the grammar will try to retrieve the ontological relations using the semantic features that are stored in the dictionary. Finally, if the previous stages fails to figure out the ontological relation between the words the context prediction will take place by predicting the identity of the proper name from the context. However, it is worth mentioning that not all of the grammar levels that are mentioned earlier are applied in this application, since that this application provides a shallow parsing only.

Lexical analysis module

This module is responsible for splitting the sentences into tokens, then matches these tokens with the dictionary in order to assign the different necessary features to each token. However, some words may be misrecognized, because of spelling mistakes or morpho-syntactic changes. For example, the most common mistake in the Arabic writings is /Hamza/ in the initial position as in "استقبل" 'receive'. The rules are able to solve this problem by investigating the morphological pattern of the wrong spelled word by the regular expression technique. For example, if a six-letters word begins with the sequence "/...است/" as in the pattern "استفعل" /?istif؟aal/, the wrong written /Hamza/ ("أ", "إ" or "آ") will be modified to "ا" according to the Arabic grammar as in the rule in (1).

(1)
```
({SHEAD|BLK|PUT|PFX},%e)(TEMP,"/(ست...(إ|أ|آ)|ي|ئ|ؤ|ت|ن|تم|نا)/",%x,^Hamza_modified)(%y,{STAIL|BLK|PUT})
:=(%e)("1>"ا,%x,X,Hamza_modified)(%y);
```

The module also deals with cases of morpho-syntactic changes as in the nominative form "علماؤ" /ʔulamaaʔu/ when it is attached to the pronoun "ه" 'its'. Rules are able to extract the deep form "علماء" 'scientists' from the surface form "علماؤه" 'its scientists' as in rule in (2).

(2) ({SHEAD|BLK|PUT|PFX},%e)("/.+(ئ|ؤ)/",^Y,%x)(POD,%w):=(%e)(%x,"ء"<"ؤ","ء"<"ئ",Y)(%w);

After the completion of the task of spelling correction, if some words are still undefined, the feature 'TEMP' will be assigned to it. Then, it will be considered as a proper name and the 'PPN' feature will be assigned to it instead of the feature 'TEMP'.

After the lexical analysis module, the semantic relations between the words of the sentences using the UNL ontological relations should be established. These relations are stated in table I below:

TABLE I

ONTOLOGICAL RELATIONS IN THE UNL SYSTEM

Tag	Relation	Definition	Example
ant	opposition or concession	Used to indicate that two entities do not share the same meaning or reference. Also used to indicate concession.	John is not Peter = ant(Peter; John)
cnt	content or theme	The object of an stative or experiental verb, or the theme of an entity.	Book about linguistics = cnt(book; linguistics)
icl	hyponymy, is a kind of	Used to refer to a subclass of a class.	Dogs are mammals = icl(mammal; dogs)
iof	is an instance of	Used to refer to an instance or individual element of a class.	John is a human being = iof(human being; John)
nam	name	The name of an entity.	The city of New York = nam(city; New York)
pof	is part of	Used to refer to a part of a whole.	John is part of the family = pof(family; John)
fld	field	Used to indicate the semantic domain of an entity.	sentence (linguistics) = fld(sentence; linguistics)

The following sections discuss the followed techniques to build the ontological relations. Each of the following sub-sections represents an attempt to recognize the identity of the proper names that have occurred in the instances; if one attempt fails to reach the recognition, the following attempt will take place.

Pattern matching

Sometimes the ontological relations mentioned in table Iwould have fixed structures with keywords that are stated in them, these structures would represent the type of the relation as in table II.

TABLE II

RELATIONS KEYWORDS AND EXAMPLES FOR THE ONTOLOGICAL RELATIONS

Relation Tag	Relation key words	Example
ant	عكس – مقابل - يقابل	الحق مقابل الباطل
cnt	عن - حول	كتاب عن التاريخ
icl	نوع / صنف من – أحد أنواع / واحد من –أصناف	القطط نوع من الحيوانات
iof	مثال ل/على	الإسكندرية مثال لبلدان مصر
nam	تسمى ب – اسم - تحت اسم	محمد اسم انسان
pof	جزء من – أحد أجزاء	الإسكندرية جزء من مصر
fld	في مجال – في علم	المورفولوجيا في علم اللغويات

The pattern matching module is responsible for building the ontological relations for structures that have keywords such as those mentioned in table (II). For example, the rule in (3) states that if a noun is followed by "جزء" 'part' then "من" 'of' and is followed by another noun, then both nouns would be linked with a 'pof' relation. However, not all of the relation keywords that are stated in table (II) are found in the corpus, but they are taken into consideration in order to achieve grammar robustness.

(3) (%a,N)("جزء",%b)("من",%b)(N,%d):=(pof(%d,with_rel;%a,rel=pof)) #L(%a,rel=pof,#CLONE;e);

Most sentences of the corpus did not contain keywords that represent the relations. Therefore, the grammar depends on the features assigned to the words that need to be related. It specially depends on the semantic classification in order to determine which ontological relation should be used. Table III lists some semantic features that have been observed:

TABLE III

SEMANTIC FEATURES OBSERVED FOR THE UW1 OF THE ONTOLOGICAL RELATIONS IN THE CORPUS

Semantic feature	Explanation	Example
HUM	person (Nouns denoting people.)	طبيب
GRO	group (Nouns denoting groupings of people or objects.)	جامعة
ARF	artifact (Nouns denoting man-made objects.)	شركة
NOB	natural object (Nouns denoting natural objects (not man-made).)	بحر
CGN	cognitive noun (Nouns denoting cognitive processes and contents.)	قانون
LCT	location (Nouns denoting spatial position.)	مدينة

For example, in the sentence "أدونيس شاعر" 'Adonis is a poet', the two words are defined in the dictionary as [أدونيسPOS=PPN, GEN=MCL, SEM=HUM] and [شاعر POS=N, GEN=MCL, SEM=HUM]. A rule can link between "أدونيس" 'Adonis' and "شاعر" 'poet' with an 'iof' relation through depending on the 'HUM' (human)feature. The rule in (4) states that if a noun such as "شاعر" 'poet' with the semantic feature 'HUM' comes after a proper noun such as "أدونيس"'Adonis', then both nouns would be linked with a 'iof' relation as in Fig. 4.

(4) (%x , PPN , HUM , ^rel = iof) (%y , HUM , ^PPN , GEN = %x) ({^N | STAIL }, %q) :=
(iof(%x , +with_rel ; %y , +rel = iof) , %01) #L(%y , #CLONE , +rel = iof ; %q) ;

```
[S:1713]
    {org}
    أدونيس شاعر
    {/org}
    {unl}
        iof(A0:شاعر ,07:أدونيس)
    {/unl}
[/S]
```

Figure 4: The UNL ontological relation using semantic features technique

Context prediction

The identity of a proper name can be predicted from the context in which it occurs. For example, the proper noun "كامبردج" 'Cambridge' in "شهادة الدكتوراه من كامبردج أو السوربون أو" 'PHD from Cambridge or Sorbonne or' doesn't have an adjacent noun that has one of the semantic features mentioned in table III, but one of the adjacent nodes (words) can help in predicting the identity of the proper name "كامبردج" 'Cambridge' which is the noun "شهادة" 'certificate'. If the rules find this list of words, then the noun "جامعة" 'university' will be inserted by the rule in (5) in order to be "شهادة الدكتوراه من جامعة كامبردج" 'PHD certificate from Cambridge university' that is linked by the 'iof' relation as in Fig. 5 .

(5)(%a,{"شهادة الماجستير"|"شهادة الدكتوراه"|"التعليم العالي"},^ins):=(%a,ins)(?[جامعة],blk,INS);

```
[S:1542]
    {org}
    شهادة الدكتوراة من كامبريدج او السوربون أو
    {/org}
    {unl}
        iof(04:جامعة ,01:كامبريدج)
    {/unl}
[/S]
```

Figure 5: The UNL ontological relation using context prediction technique

Knowledge base

In the sentence "وقع رئيس أذربيجان السابق حيدر علييف" 'former President of Azerbaijan Heydar Aliyev has signed', the two nouns "رئيس" 'president' and "أذربيجان" 'Azerbaijan' cannot be linked with a direct relation, given the fact that the dictionary includes [أذربيجانPOS=PPN, GEN=MCL, SEM=LCT, CAR=ONE] and [رئيس POS=N, GEN=MCL, SEM=HUM], since "أذربيجان"'Azerbaijan' is not an instance for "رئيس" 'president'. However, such adjacent words with those semantic features should be linked with 'mod' relation, but not an ontological one as in rule (6); the mod relation is not displayedin the final output, but it is a method to block applying the ontological relation. All of the adjacent nodes; the context around the proper name "أذربيجان" 'Azerbaijan' fail to help in recognizing its identity. Therefore, only the dictionary features can help in predicting the identity of "أذربيجان" 'Azerbaijan', as it is a location 'SEM=LCT' and it is the only one in the world 'CAR=ONE', so it could be concluded that it is an instance of a country. In the case of predictionof a proper name identity, the rule inserts the identity noun "دولة" 'country' before the proper name by the insertion rule described in (7). The 'iof' relation will link "أذربيجان" and "دولة" as in Fig. 6.

```
(6)
(SHEAD,%c)(N,HUM,^with_rel,^INS,%b)(%a,N,PPN,^GEN=%b):=(mod(%b;%a,rel=mod)) #L(%c;%a,rel=mod,#CLONE);
(7)
({^GRO,^ARF,^HUM,^NOB,^LCT,^CGN|SHEAD
|LCT,PPN|PPN,GRO|CGN,DEF|HUM,PLR|"سكان"|"عاصمة"},%a)(%b,LCT,PPN,ONE,{^ins|ins,SPLIT},^with_add,^rel=iof)(
{^"دولة",^NOB,^LCT|STAIL|node_left_del|LCT,PPN|COO|PPN,GRO|DEF},%c):=(%a)(?[دولة],blk,INS)(%b,ins)(%c);
```

```
[S:2001]
    {org}
    وقع رئيس أذربيجان السابق حيدر علييف
    {/org}
    {unl}
        iof(04:دولة ,01: أذربيجان)
    {/unl}
[/S]
```

Figure 1: The UNL ontological relation using knowledge base technique

B. Tools and Engines

1) SEAN: is the acronym for Shallow Enhanced ANalyser. It is fully automatic; it does not allow for any human intervention. It is a multi-document analyzer. Moreover, it is a word-driven analyzer: the unit of analysis is a word that is provided by the user. It is also a shallow analyzer: the analysis targets the surface structure of natural language sentences.

SEAN is appropriate for information retrieval and extraction task, because it provides a rather rough and partial analysis of the natural language input. SEAN has been developed by the engineering team in the Library of Alexandria.

Dictionaries, N-rules, T-rules and D-rules tabs in SEANare provided the dictionary, normalization rules, transformation and disambiguation rules. In the Sean Documents tab, the NL documents can be uploaded either as web links or a text file in the UTF8 format.

Moreover, the Process tab, allows the user to search for a word in the uploaded text. The number of words around the searched word can be specified from the combo box 'Concordance'. The process tab consists of 4 sub-tabs; concordance, UNL corpus,

Knowledge base and trace tabs. By clicking the 'search' button under the comb box, search results will be shown in the concordance tab in the left pane as shown in Fig. 7.

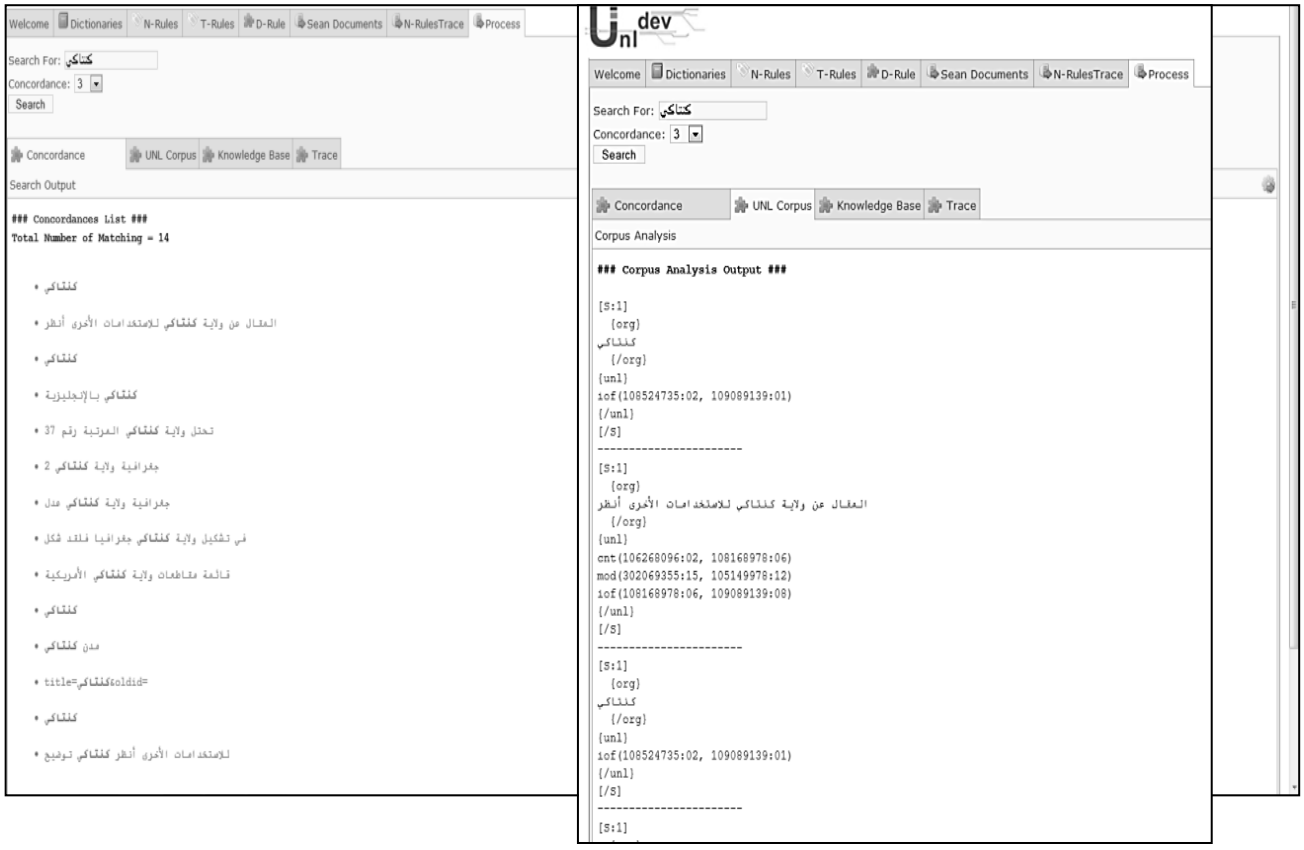

Figure 7: The "Process" tab.

The selected natural language text from the concordance result is processed by the selected dictionary and the selected rules files, the UNL expressions of all search results are shown in the sub-tab 'UNL corpus'. The behavior of the applied rules can be viewed in the sub-tab 'trace'.

2) EUGENE: This tool is responsible for generating the natural language sentences out of semantic networks represented in the UNL format. In its current release, it is a web application developed in Java and available at the UNLdev4. EUGENE is an acronym for dEp-to-sUrfaceGENErator. As a multilingual engine, EUGENE must be parameterized to the target natural languages with the following files that are provided through EUGENE's interface: The input document in the UNL document structure, i.e., the universal semantic network to be generated in natural language, the UNL-NL (generation) dictionary, i.e., a lexical database where UWs are mapped into natural language entries, along with the corresponding features, the UNL-NL (generation) transformation grammar, i.e., a set of transformation rules used to convert the UNL graphs into natural language sentences and the UNL-NL (generation) disambiguation grammar, i.e., a set of disambiguation rules used to improve the results of the tokenization and of the transformation[5][31].

5 KEYS (KEY'S INTERFACE)

KEYS is an automatic language-independent knowledge extraction system, it automatically extracts structured information, from unstructured machine-readable natural language documents. The system is able to work with any language as long as it contains the required resources of this given language. The following sub-sections will describe how KEYS works starting from the point of files uploading to the point of obtaining the results.

[4]http://dev.undlfoundation.org/index.jsp
[5]http://www.unlweb.net/wiki/EUGENE

A. Uploading documents

The user has to select the language of the file that will be uploaded from a dropdown list. The user can also select the desired text file or zip folder from his file system by clicking on the browse button then clicking on the upload button. The user can add a URL file, by inserting the URL.

B. Search for a query

The user has to select the desired language, concordance size and documents in order to search in these documents, then enter the search query and click the search button. The results will be viewed in tabs (Visualization, Simplified and Eugene).

1) Visualization: The output will be presented in the form of a graph by clicking on the "Visualization" tab. The output of this option depends on the results provided through SEAN. In this view, the user can click on each node in the graph to display its relations; the thick arrow refers to the most frequent instance of the word as shown in Fig. 8.

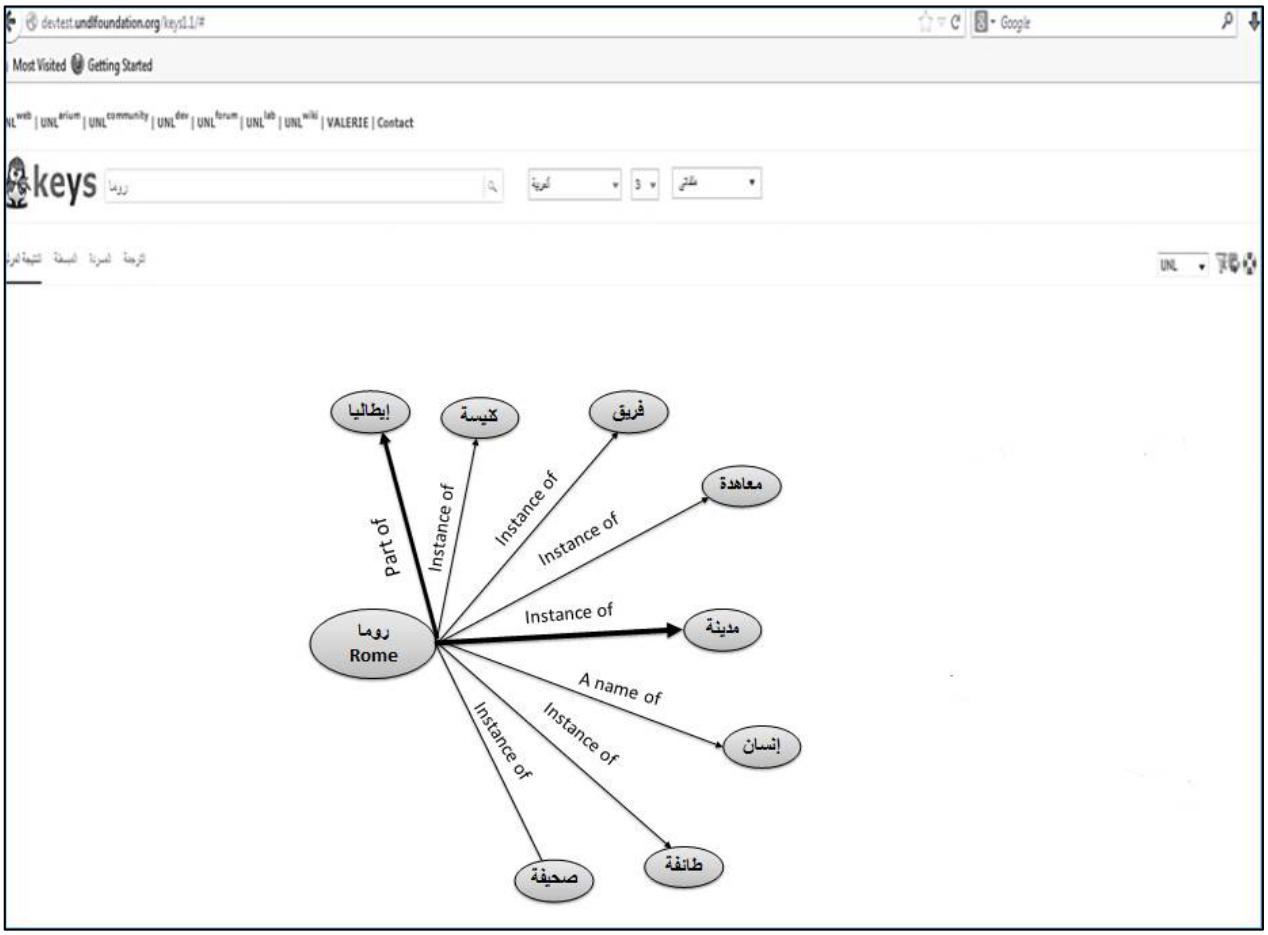

Figure 8: KEYS output in the UNL view (visualization)

2) *Simplified KB:* The output will be presented in the form of UNL expressions by clicking on the "Simplified" tab. The output of this option also depends on the results provided through SEAN. The output will appear as shown in Fig. 9.

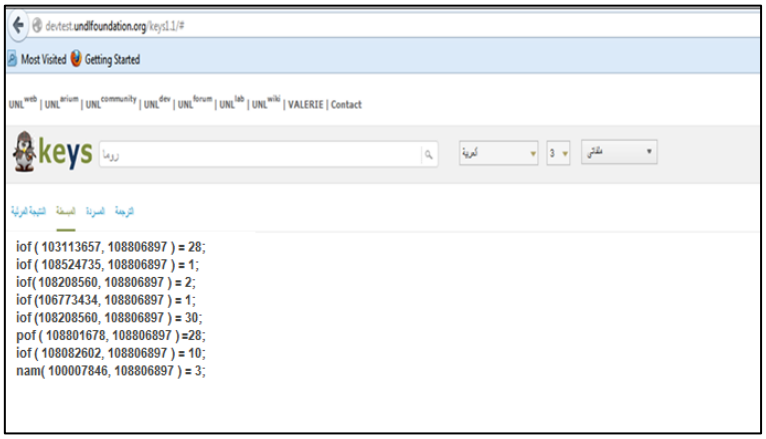

Figure 9: KEYS output in the UNL view (simplified)

3) *EUGENE:* The output will be presented in the form of natural language sentences by clicking on the "EUGENE" tab. The output of this option depends on the results of the generation tool EUGENE. The output will appear as shown in Fig. 10.

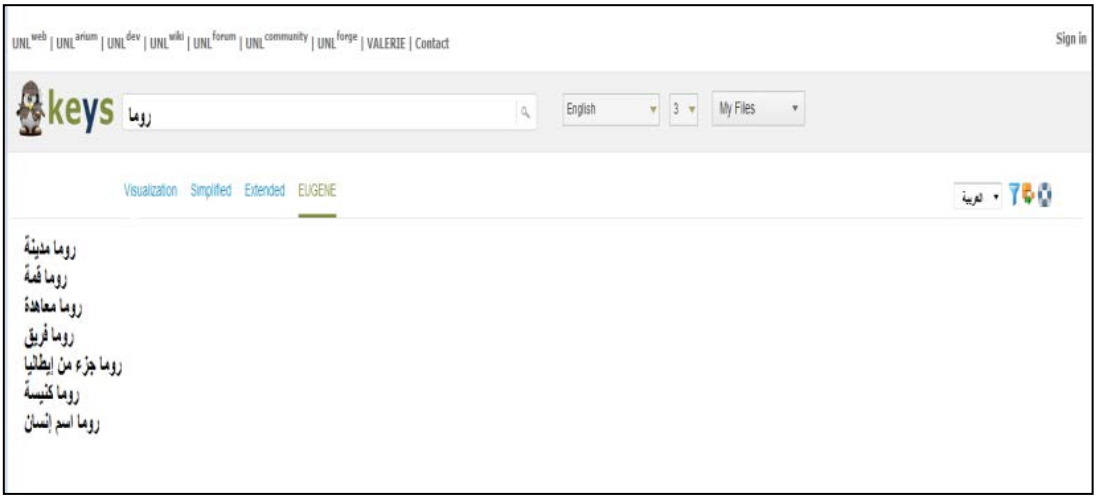

Figure 10: KEYS output (Eugene)

6 EVALUATING THE RESULTS

Evaluation has been performed in order to investigate the accuracy and robustness of the grammar. The used data consists of 1000 proper nouns as keys for search with total number of occurrences being 22,000. The instances are divided into a training set which includes 17,000 instances and a testing set which includes 5,000 instances. The same proper name used in the training data has been tested in different contexts, different from the trained instances. For example, the output of the proper name "هدسون" 'Hudson' in the trained data was (14 instances), while the tested contexts represent 4 instances.The primary scores are precision and recall. Let N_{key} be the total number of filled slots in the answer key, $N_{response}$ be the total number of filled slots in the system response, and $N_{correct}$ be the number of correctly filled slots in the system response (i.e., the number which match the answer key). Then

$$\textbf{Precision} = \frac{N\ correct}{N\ response} = 19.500/21.000 = 0.92$$

$$\textbf{Recall} = \frac{N\ correct}{N\ \ keys} = 19.500/22.000 = 0.886$$

The F measure was calculated with the equation: F= (2 × precision × recall)/(precision + recall) and the accuracy was 90.2 %. These equations were calculated for each answer key in the corpus, then all of their results were added to provide the total accuracy of the corpus.

In addition, a different set of proper names other than those used in the training and testing sets have been used to see whether the system has enough knowledge to search for any other entities in other contexts. For instance, Fig. 11, 12, and 13 represent the output samples of the proper names "الإسكندرية" 'Alexandria', "هونج كونج" 'Hong Kong' and "مونتريال" 'Montreal' respectively which were not included in the 1000 proper names we worked with. Fig. 11, 12 and 13 reflect that the system has learned abstracted knowledge that made it able to deal with both new keys and new contexts.

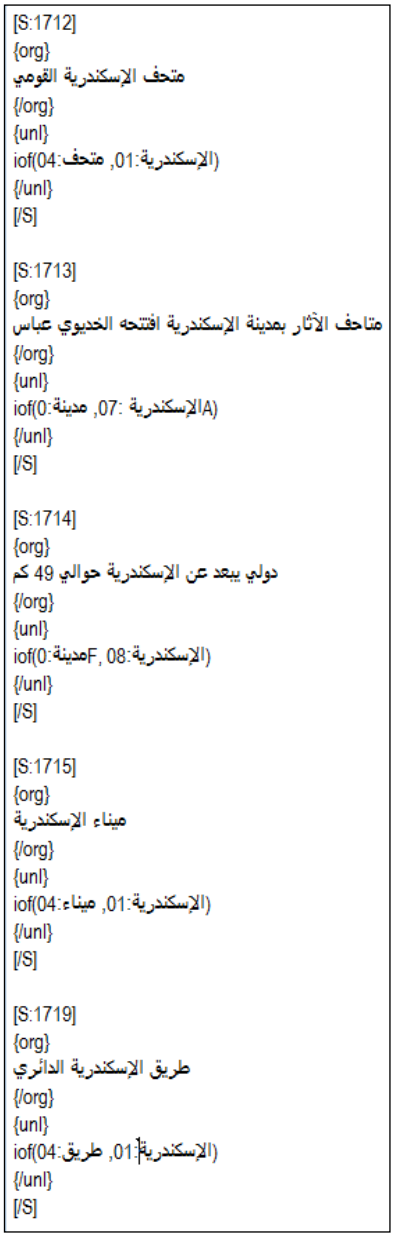

Figure 11: Sample of "الإسكندرية" output

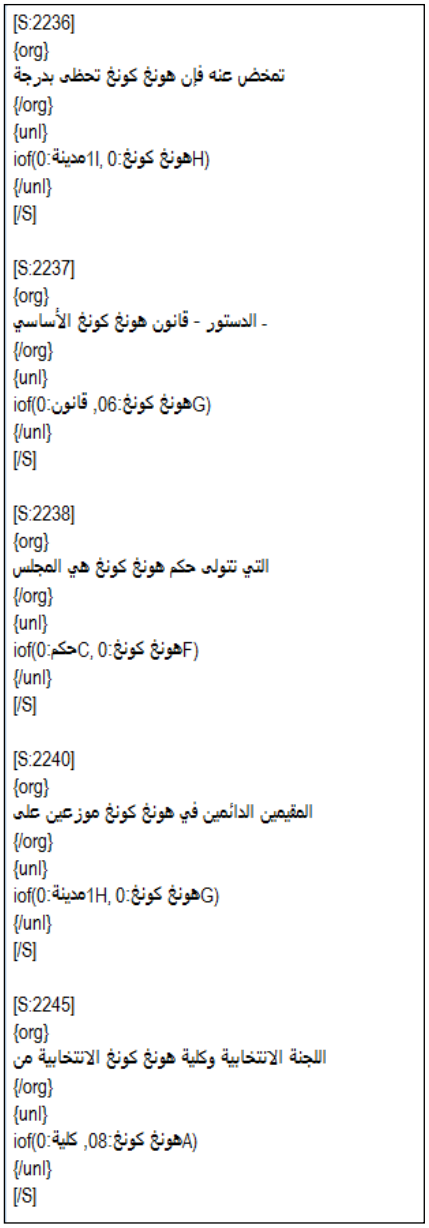

Figure 12: Sample of " هونغ كونج"output

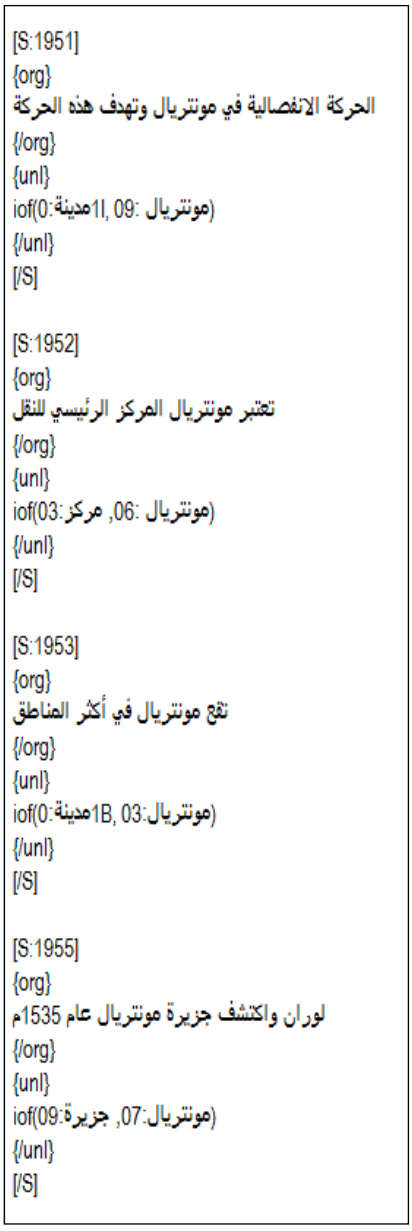

Figure 13: Sample of "مونتريال"output

7 CONCLUSIONS

Many applications depend on the automatic extraction of structure data from unstructured data for better means of querying, organizing, and analyzing data. KEYS is a knowledge extraction system that promises to fulfil the human needs in providing an

easy access to the vast amount of information that is readily available on the internet. The amount of information on the internet is rapidly increasing, it is increasing every second, which makes benefiting from this amount of information difficult. Hence, the importance of knowledge extraction systems is manifested in providing an easy method to obtain the needed information. Knowledge extraction systems maximize the magnitude of utilizing the available information. In this article, the infrastructure of KEYS system is discussed. The linguistic resources and the tools involved in KEYS are presented, they are all provided in an open-source form for free at www.unlweb.net. The precision measurement of the Arabic grammar was 0.92 while recall measurement was 0.886.

REFERENCES

[1] J. Hobbs, and E. Riloff, E. *Information Extraction*, *Handbook of Natural Language Processing, 2nd Edition*, Editors: Nitin Indurkhya and Fred J. Damerau, Chapman & Hall/CRC Press, Taylor & Francis Group, 2010.
[2] S. Patwardhan ,"Widening the field of view of information extraction through sentential event recognition", A dissertation submitted to the faculty of The University of Utah in partial fulfilment of the requirements for the degree of Doctor of Philosophy 2010.
[3] W. Lehnert, and J. Cowie, "Information Extraction". Communications of the ACM 39, 80–91, 1996.
[4] G. DeJong, "Prediction and Substantiation: A New Approach to Natural Language Processing "*A Multidisciplinary Journal 3*, 251–271. 1,1996.
[5] G. DeJong, "An Overview of the FRUMP System". *In Strategies for Natural Language Processing*, W. Lehnert and M. Ringle, Eds. Erlbaum, Hillsdale, NJ, 1982, pp. 149–176.
[6] R. Schank, and R. Abelson, "Scripts, Plans, and Understanding". Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.
[7] R. Cullingford, "Computer Understanding of Newspaper Stories", PhD thesis, Yale University, 1978.
[8] G. Silva, , and D. Dwiggins, , "Towards a Prolog Text Grammar", ACM SIGART Bulletin 73, 20–25, 1980.
[9] J. Cowie, "Automatic Analysis of Descriptive Texts". *In Proceedings of the First Conference on Applied Natural Language Processing*, Santa Monica, CA, pp. 117–123, 1983.
[10]N. Sager, "Natural Language Information Processing: A Computer Grammar of English and Its Applications", Addison-Wesley, Boston, MA, 1981.
[11] G. Zarri, , "Automatic Representation of the Semantic Relationships Corresponding to a French Surface Expression", *In Proceedings of the First Conference on Applied Natural Language Processing* ,Santa Monica, CA, pp. 143–147, 1983.
[12] R. Grishman, , and B. Sundheim, "Message Understanding Conference - 6: A Brief History". *In Proceedings of the 16th International Conference on Computational Linguistics* , Copenhagen, Denmark, pp. 466–471, 1996.
[13] R. Merchant, "TIPSTER Program Overview", *In TIPSTER Text Program Phase I*: Proceedings of the Workshop (Fredricksburg, VA), pp. 1–2, 1993.
[14] P. Altomari, and P. Currier, "Focus of TIPSTER Phases I and II". In TIPSTER Text Program Phase II: Proceedings of the Workshop , Vienna, VA, pp. 9–11, 1996.
[15] F. Ruth Gee, "The TIPSTER Text Program Overview", In Proceedings of the TIPSTER Text Program Phase III , Baltimore, MD, pp. 3–5, 1998.
[16] G. Krupka, L. Iwariska, P. Jacobs, and L. Rau, "GE NLToolset: MUC-3 Test Results and Analysis", *In Proceedings of the Third Message Understanding Conference (MUC-3)* ,San Diego, CA, pp. 60–68, 1991.
[17] J. Hobbs, "SRI International's TACITUS System: MUC-3 Test Results and Analysis", *In Proceedings of the Third Message Understanding Conference (MUC-3)* , San Diego, CA , May 1991, pp. 105–107.
[18] W. Lehnert, , C. Cardie, D. Fisher, E. Riloff, and R. Williams, "MUC-3 Test Results and Analysis", *In Proceedings of the Third Message Understanding Conference (MUC-3)* , San Diego, CA, May 1991 , pp. 116–119.
[19] R. Grishman, J. Sterling, , and C. MacLeod, "MUC-3 Test Results and Analysis" , *In Proceedings of the Third Message Understanding Conference (MUC-3)* ,San Diego, CA, pp. 95–98, 1991.
[20] M. Banko , O. Etzioni, "Strategies for Lifelong Knowledge Extraction from the Web", Turing Center, University of Washington Computer Science and Engineering, 2007.
[21] L. Elikuil, "Information Extraction from the World Wide Web: A Survey". Norwegian Computer Center, Report no. 945, 1999.
[22] N. Naw and E. Hlaing, "Relevant Words Mining with Compiling Technique", *International Journal of Emerging Technology and Advanced Engineering* , Volume 4, Issue 4, 2014.
[23] N. Naw and E. Hlaing, "Relevant Words Extraction Method for Web Recommender System", *International Conference on Advances in Engineering and Technology (ICAET'2014)* March 29-30, 2014.
[24] R. Grishman, "Information Extraction: Techniques and Challenges". *International Summer School on Information Extraction (SCIE'97)* Frascati, Italy (=Lecture Notes in Computer Science,1299),10-27.Berlin:Springer. 1997.
[25] S. Alansary, M. Nagi, N. Adly, "UNL+3: The Gateway to a Fully Operational UNL System", *in Proceeding of 10th Conference on Language Engineering*, Cairo, Egypt, 2010.

[26] H. Uchida, "UNL: Universal Networking Language – An Electronic Language for Communication, Understanding, and Collaboration". UNU/IAS/UNL Center. Tokyo, Japan, 1996.
[27] H. Uchida, M. Zhu, "The Universal Networking Language beyond Machine Translation", UNL Foundation, 2001.
[28] C. Jesús, A. Gelbukh, E. Tovar (eds.), *"Universal Networking Language: advances in theory and applications"*. Mexico City: National Polytechnic Institute, 2005.
[29] H. Uchida, M. Zhu, "UNL2005 for Providing Knowledge Infrastructure", *in Proceeding of the Semantic Computing Workshop (SeC2005)*, Chiba, Japan, 2005.
[30] S. Alansary, "MUHIT: A Multilingual Harmonized Dictionary", The 9th edition of the Language Resources and Evaluation Conference, Reykjavik, Iceland, 26-31 May, 2014.
[31] S. Alansary, M. Nagi and N. Adly, "Generating Arabic Text: the Decoding Component in an Interlingual System for Man-Machine Communication in Natural Language", 6th International Conference on Language Engineering, Cairo, Egypt, December 2006.

BIOGRAPHY

Dr. Sameh Alansary

Director of Arabic Computational Linguistics Center Bibliotheca Alexandrina.

He is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars. He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now. Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He Has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

Magdy Nagi

He is a senior consultant, ICT Sector, Bibliotheca Alexandrina.

He is a Professor in the Computer and Systems Engineering department, Faculty of Engineering, Alexandria University. He obtained his Ph.D. from the University of Karlsruhe, in 1974, where he served as Lecturer for two years and as a Consultant to its Computer Center from 1974-1990. During this period he also served as Consultant to many companies in Germany such as Dr. Otker, Bayer, SYDAT AG, and BEC. He served, since 1995, as Consultant to the Bibliotheca Alexandrina. Among his activities were the design and installation of Bibliotheca Alexandrina's network and information system, namely a trilingual information system that offers full library automation. In 2001, he got appointed as the Head of the Information and Communication Technology (ICT) Sector of the Bibliotheca Alexandrina and occupied that post till 2012. He currently serves as a senior Consultant to the ICT Sector and continues to oversee the various projects and partnerships established between the ICT Sector and many international institutions. Dr. Nagi is a member of the ACM and the IEEE Computer Society as well as several other scientific organizations. His main research interests are in operating systems and database systems. He is author/co-author of more than 100 papers.

KEYS: نظام استخلاص معرفي اعتمادا على البنية التحتية المعرفية للغة الشبكات العالمية

سامح الأنصاري[1*] ، مجدي ناجي[2**]

مكتبة الإسكندرية، الشاطبي، الإسكندرية، مصر
*قسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية، الشاطبي، الإسكندرية، مصر
[1]sameh.alansary@bibalex.org
** قسم هندسة النظم والحاسب، كلية الهندسة، جامعة الإسكندرية، الإسكندرية، مصر
[2]magdy.nagi@bibalex.org

ملخص

في وجود ثورة المعلومات واتاحة كم كبير منها على صفحات الانترنت زاد احتياج الانسان لاستخلاص معلومات محددة من هذه الصفحات ، لذلك فإن هدف هذه الورقة البحثية تقديم النظام (KEYS) كنظام استخلاص معرفي يهدف إلى استخلاص واسترجاع المعلومات إذ يقوم بالبحث عن المعلومات داخل نصوص ممثلة دلاليا باستخدام لغة الشبكات العالمية (UNL) مما يجعل مهمة استخلاص واسترجاع المعلومات أكثر دقة وعندها يكون من المتوقع أن تكون النتائج ذات جودة عالية ؛ إذ يتم تحليل النص المصدر تحليلا سطحيا وتحويله إلى شبكة دلالية تحتوي على علاقات انطولوجية محددة ممثلة باستخدام لغة الشبكات العالمية، بعد ذلك يتم توليد آلي لأي لغة طبيعية من هذه الشبكة الدلالية. وبهذا يكون من المتوقع أن يقدم هذا النظام نهجا جديدا لتحديد الكيان الاسمي وهو استخراج الاسماء بجميع تصنيفاتها الانطولوجية من اللغة الطبيعية أيا كانت هذه اللغة.