Software Engineering 2012, 2(2): 36-42
DOI: 10.5923/j.se.20120202.04

Towards Arabic Noun Phrase Extractor (ANPE) Using Information Retrieval Techniques

Islah K. Gharaibeh[1,*], Natheer K. Gharaibeh[2]

[1]Prince Abdullah Bin Ghazi Faculty of IT Al-Balqa Applied University
[2]Ajlun College ,Al-Balqa Applied University
islahgh@yahoo.com, natheer_garaybih@yahoo.com

Abstract Information Retrieval system aims to help people find relevant information when they request it. This paper is interested in Noun Phrases (NPs) for Arabic language; generally the Noun Phrase Extractor is made up of three modules: tokenization; part-of-speech tagging; noun phrase identification. Also we will preview Noun Phrase Extractors, compare between them according to their suitability for building the new system, We focused on the effect of using the Arabic noun phrases that has been extracted on the information retrieval system in the light of computing measures of evaluation (Recall and Precision), we found that we get more relevant documents from the retrieved ones when using NPs rather than using single terms, on the other side number of retrieved documents will be decreased.

Keywords Information Retrieval (IR) , Natural Language Processing (NLP) , World Wide Web (WWW) , Arabic Noun Phrase Extractor (ANPE) , Noun Phrase (NP), Tokenization, Tagging, Parsing, Recall, Precision

1. Introduction

In the beginning of the 1990's, the WWW becoming a universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before[1], however the introduction of the WWW put the Information Retrieval at the center of the stage.

Too many researchers talked about Information Retrieval Systems, and have introduced definitions such as:

Salton[2] defined it "Information-retrieval systems process files of records and requests for information, and identify and retrieve from the files certain records in response to the information requests. The retrieval of particular records depends on the similarity between the records and the queries, which in turn is measured by comparing the values of certain attributes to records and information requests."

Kowalski[3] defined it "An Information Retrieval System is a system that is capable of storage, or maintenance information. Information in this context can be composed of text (including numeric and data), images, audio, video, and other multi-media objects)". We can say that the Primary goal of an IR system is : "Retrieve all the documents which are relevant to a user query, while retrieving as few non-relevant documents as possible."[4]

Electronic age.

Arabic Information Retrieval (Arabic IR) has recently become a focus of research and commercial development due to the vital necessity of such tool for people in the The number of Arab-speaking Internet users in 2002 was about 4.4 million, about 1.5% of the Arab world population. But on the other side of the picture few search engines are available to Arabic-speaking users, even though efforts are in progress to serve the increasing number of users[21].

Arabic is a challenging language for information retrieval (IR) for a number of reasons some of them mentioned in 3.2, Those problems make exact keyword match inadequate for Arabic retrieval[22].

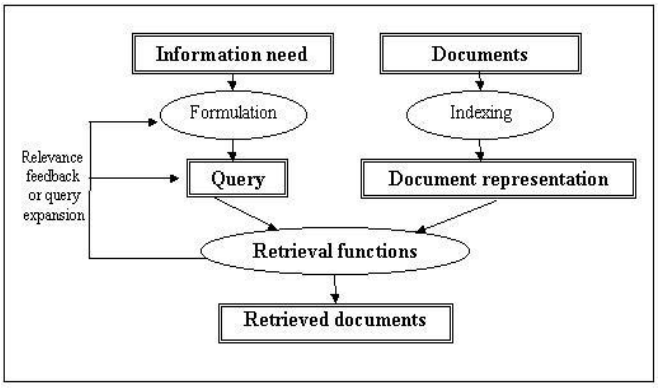


Figure.1. The Retrieval Process[4]

One of the major limitations of the Arabic IR system developer is **the lack of adequate resources** that could help test their system to get good evaluation of the system's performance in the real world, these resources are[21]:

- **Corpora**: the only large scale resources known and

* Corresponding author:
islahgh@yahoo.com (Islah K. Gharaibeh)
Published online at http://journal.sapub.org/se
Copyright © 2012 Scientific & Academic Publishing. All Rights Reserved

available to users are the LDC collection from Agency France Press (AFP) and the Al-Hayat newspaper collection from the European Language Resources Distribution Agency.

- Lexicons: such as monolingual and bi-lingual dictionaries, available on line are the Ajeeb dictionary and the Ectaco, which were used in some IR experiments.

- Tools: Arabic has a high degree of inflection; its morphology is a challenging task for IR systems, to solve this problem two major approaches were used to build morphological analysers, Rule based morphological analyser by Shereen Khoja (2001), and the Finite state Transducer of Xerox by Beesely (1996, 1998).

1.1. Noun Phrase (NP) Definition

In general, phrases are defined as pairs of words (or longer sequences) are treated as single terms. They are frequently used in languages, many researchers in the Information Retrieval field have mentioned them and made their researches on extracting and indexing them since the 1960s.

Salton[2] did some exploratory indexing using noun phrases. He used syntactic phrases in comparison to statistically chosen phrases, he found that statistical phrases gave somewhat better retrieval results. He defined "term phrases" as phrases consisting of sequences of related text words. Phrases carry a more specific meaning than single terms included in the phrases, for example "Computer Science" or "Computer program" is more specific than "Computer", and he has made a procedure for statistical phrase indexing at 1975.

Fagan[20] used both "statistical" and "syntactic" phrases, a statistical phrase is defined by constraints on the number of cooccurrences and cooccurrences of its component word and/or the proximity between occurrences of components in documents, a syntactic phrase may be characterized by some of the same criteria as a statistical phrase, but in addition must obey some constraints on the syntactic relationships among its component words. He said that the objective of using phrases as content indicators is to take advantage of the fact that phrases identify concepts that are more specific than the concepts identified by their components in isolation, Fagan extended the work done by Salton.

El-Naggar[11] defined a phrase structure grammar of the Arabic Language, he classified the phrases modifiers into two categories :

The first one contains noun phrase post modifiers, in which the ordered sequence of these modifiers is significant; this is the key to construct these noun phrase structures. These modifiers are: noun complement (NC), Adjective (ADJ), apposition (APP) and correlation (COR)

The second consists mainly of Verb Phrase modifiers like adverbs[11]. More about El-Naggar's research mentioned in the next section.

1.2. Why do we need to Extract NPs

It has been assumed by researchers that in text it is the noun phrases that are the content-bearing elements[5], as well as some of them called the noun phrase extraction as concept extraction[8], certainly NPs are more content-bearing than single words but phrases are not a full representation of meaning, yet NPs are good indicators of text content, and for traditional IR, that is what we want.

Smeaton[5] illustrates that When we perform indexing by phrases we index into a vocabulary, the set of phrases, which is richer than the set of words or word senses, thus if we have a richer representation format, and we can translate text into this accurately, we should get better quality retrieval.

Also Haddad[19] said that previous studies have shown that the use of phrases to represent a document's content can enhance the effectiveness of an automatic Information Retrieval (IR) system. He added that phrases, and specially noun phrases (NPs), have been proposed as more sophisticated representation.

2. Literature Review

Accordingly researchers have worked hardly in the area of extracting NPs, not only in the English language but also others, approximately, most of NP extractors (sometimes called phrasers or NP parsers) include the three main steps : tokenization, tagging and forming noun phrases (these will be explained also in our system at section4).

- FastNPE (Fast Noun Phrase Extractor): Developed by collaborators in the AI Lab in the department of MIS at the University of Arizona[Alleviating]. In order to extract terms from a text document, it operates by using a stopword list and a stemlist, combining these lists, and comparing the final list of words with all of the words found in the document.

- NPtool: Commercially available from Lingsoft in Finland[Building]. It has two main operations. In the first, each term of the text is given a context-free description and part-of-speech tag, and in the second, the actual noun phrases are output[8].

- Chopper: Developed by Ken Haase at MIT, will parse a text by breaking it down into constituent sentences or phrases. The input document can be sent to the parser through MIT´s World-Wide Web page. The output text is received back in sentence-reverse order, each term of the text also having been tagged with its part-of-speech tag[8].

- AZ Phraser: By the MIS group at the University of Arizona[10]. In order to improve the FastNPE, it performs three tasks. Tokenization, tagging words and finally output a list of noun phrases[8].

Figure.2 Shows the quality evaluation for the previous four NP extractors by calculating the recall and precision measures taken from[7].

	FastNPE	NPtool	Chopper	AZ Phraser
Recall	50%	95%	97%	92%
Precision	80%	96%	90%	86%

Figure.2. Recall and precision results for the four parsers

- Phrase Structure Grammar of the Arabic Language by Ayman El-Naggar[11]: It proposed eight structures covering all the categories of noun phrases. Five of the structures given in table.3, they consist mainly of Noun (N) as the basic unit and its maximum allowable post-modifiers which could follow it. These post-modifiers have a specific sequence which is the key to construct these structures and according to the number of these post-modifiers, the structures vary The NP structures as well as some of the ARABIC categories they constitute.

Table 3. Five structures of NPs proposed by El-Naggar

CATEGORY	NP STRUCTURES
OBJECT	NP1 = (COR) (APP) (ADJ) (NC) N
ADJECTIVE	NP2 = (NC) N
'MAFOOL MO TLAK' (CASE 1)	NP3 = N
'MAFOOL MO TLAK' (CASE 2)	NP4 =[ADJ] N
CORROLA TION	NP5 =[NC] N

- The phrase generation system by Gerard Salton[2] **:** The phrase formation system used builds two-term phrases by combining the head of a constituent with the head of each constituent that modifies it[2] . So he made the identification of nominal constructions, then the assignment of importance weights to the term phrases, and the choice of phrases as indexing units.

- Multilingual Finite-state Noun Phrase Extractor by Anne Schiller[10]**:** Presents a tool for noun phrase mark-up based on finite state techniques and statistical part-of-speech disambiguation, the overall architecture is language -independent and can be adapted for multiple languages (Dutch, English, French, German, Italian, Portuguese, Spanish).

- French Noun Phrase Indexing and mining by Hatem Haddad[19]**:** The experiment was conducted using two French test corpora, showing that combining noun phrase indexing with associative relations can improve the information retrieval system performance, especially at low recall.

- PhraseX and the SPECIALIST Minimal Commitment Parser[15] : A program that extracts noun phrase strings by referring to the syntactic structure provided by the SPECIALIST minimal commitment parser. The output produced is in the tradition of partial parsing[18] and concentrates on the simple noun phrase, what Weischedel at[18] call the "core noun phrase" that is, a noun phrase with no modification to the right of the head.

3. Arabic Language Structure

Arabic continues as it has done for the last 14 centuries to ensure the linguistic unity of over 200 million speakers spread over 20 countries which stretch from Morocco to Iraq. Arabic is the language of the Quran (the sacred book of Islam) as stated in its Surat (verse) 42:4, "... thus we have inspired unto you an Arabic Quran."

Throughout history, Muslims have believed that the purity, beauty and eternity of Arabic as a medium of expression derive from the Quran. Thus the eternity of the Quran confers eternity on the Arabic language[14]. It is this kind of attitude and belief which have given Arabic its intrinsic character as a language which has barely changed over many centuries. The structure of Arabic has been preserved throughout the Arabic-speaking world due mainly to the mystic power of the Quran[14].

The modern form of Arabic is called Modern Standard Arabic (MSA) and it is the form used by all Arabic-speaking countries in publications, the media and academic institutions. MSA is spoken by people from different Arab countries where the local dialect may not be mutually intelligible. MSA is a simplified form of Classical Arabic, and follows its grammar. The main differences between Classical and MSA are that MSA has a larger (more modern) vocabulary, and does not use some of the more complicated forms of grammar found in Classical Arabic[12].

3.1. Historical Survey of Arabic Grammar

The grammar of Arabic was investigated and then developed in the seventh century by Al-Duali who is widely regarded as having set the foundation stone of Arabic grammar, followed later by Al-Khalil and Sibawayh[14].

The grammar of present day written Arabic is still based essentially on the traditional three parts of speech:

(ism = the noun, fi3l = the verb, and harf = the particle) as laid down by Al- Duali and then later by Sibawayah who wrote Al-Kitab (The Book) prescribing the structural rules and conventions of Arabic[14].

However, it may true to say that Arab grammarians were not so much concerned with the description and functions of Arabic linguistic features as such. They were merely prescribing the language for their contemporaries - how Arabic should ideally be written and spoken. Thus Al-Nahw al-salih or Al wadih (the right way) established the foundations of the language system. The purpose was to prevent the misuse or abuse of the language of the Quran during the Islamic conquests where many languages were replaced by Arabic and their peoples Islamized and Arabized[14].

3.2. Challenges of Arabic in NLP and IR

Arabic is the official language of twenty Middle East and African countries, and is the religious language of all Muslims, regardless of their origin. It is therefore surprising that very little work has been done on Arabic corpus linguistics[12].

Arabic differs from Indo-European languages syntactically, morphologically and semantically. It is a Semitic language whose main characteristic feature is that most words are built up from roots by following certain fixed patterns and adding infixes, prefixes and suffixes. It is an old language, and what is now known as Classical Arabic was standardized around fourteen centuries ago[12].

Parsing Arabic sentences is a difficult task. The difficulty comes from several sources[13]:

(1) The length of the sentence and the complex Arabic syntax,
(2) The omission of diacritics (vowels) in written Arabic "altashkiil",
(3) The free word order nature of Arabic sentence, and
(4) The presence of an elliptic personal pronoun "alDamiir almustatir".

Despite the challenges mentioned above (and others) there are some efforts in Arabic analysis have been made in recent years but it still insignificant compared with what have done in the other languages.

3.3. Arabic Tagset

Since the grammar of Arabic has been standardized for centuries, the initial tagset is derived as in[12] from this grammatical tradition rather than from an Indo-European based tagset. The reason for this is that Arabic is a very different language from Indo-European languages – as discussed above and should have its own tagset. Also, Arabic linguists will be basing their studies on a traditional Arabic grammar rather than an Indo-European grammar.

Shereen Khoja[12] illustrates that Arabic grammarians traditionally analyse all Arabic words into three main parts-of-speech – as illustrated also above in section 3.1 . These parts-of-speech are further sub-categorized into more detailed parts-of-speech which collectively cover the whole of the Arabic language (see also[9]), (for more details look at Figure.6).

The three main parts-of-speech are:

1. Noun: -A noun in Arabic is a name or a word that describes a person, thing, or idea. These nouns could be further sub-categorised by number, gender and case. The next section will explain more about nouns in Arabic.

2. Verb:--The verb classification in Arabic is similar to that in English, although the tenses and aspects are different. The Verb tag can be sub-categorised into Perfect, Imperfect, and Imperative. Further sub-categorisation of the Verb class is possible using number, person and gender.

3. Particle:-The Particle class includes: Prepositions, Adverbs, Conjunctions, Interrogative Particles, Exceptions, and Interjections.

3.4. Nouns in Arabic Language

A noun in Arabic is a word that indicates a meaning by itself without being connected with the notion of time[17] . There are two main kinds of noun: variable and invariable.

Variable nouns have different forms for the singular, the dual, the plural, the diminutive, and the relative.

Variable nouns are again divided into two kinds: inert and derived.

1. Inert nouns (Also called Primitives[12]) are not derived from another word, i.e. it does not refer to a verbal root. Inert nouns are divided into two kinds: Concrete nouns (e.g., lion), and Abstract nouns (e.g., love).

2. Derived nouns (Also called Derivatives[12]) are taken from another word (nouns derived from verbs, nouns derived from other nouns, and nouns derived from particles[12]) (e.g. office); they have a root to refer to. A derived noun is usually close to its root in meaning. It indicates, besides the meaning, the concrete thing that caused its formation (case of the agent-noun), or underwent its action (case of the patient-noun), or any other notions of time, place, or instrument.

The following are the noun types[17] :

A **genus noun** indicates what is common to every element of the genus without being specific to any one of them. It is the word naming a person, an animal, a thing or an idea.

Example: رجل man كتاب book

An **agent noun** is a derived noun indicating the actor of the verb or its behaviour. It has several patterns according to its root.

Example: دارس the person who studies

A **patient noun** is a derived noun indicating the person or thing that undergoes the action of the verb. Patient nouns have several patterns depending in the verbal root. Example: مدروس the thing that has been studied

An **instrument noun** is a noun indicating the tool of an action. Some instruments are derived; some are inert. Example: مفتاح key

An **adjective** is considered to be a type of noun in traditional Arabic grammar. It describes the state of the modified noun. Example: جميل beautiful سيد Mr. محاضر Professor كبير big

An **adverb** is a noun that is not derived and that indicates the place or the time of the action. Example: شهر Month مدينة city شمال north

A **proper noun** is the name of a specific person, place, organization, thing, idea, event, date, time, or other entity. Some of them are solid (inert) nouns some of them are derived.

3.5. Sentences in Arabic Language

The grammar specifies the structure of the Arabic sentence. The Arabic sentence is generally classified as either nominal sentence or verbal sentence[Chart]. Accordingly, a nominal sentence is defined as a sentence which begins with a noun, whereas a verbal sentence is one which begins with a verb[16].

In each case the sentence is either simple or compound. The difference between the simple sentence and the compound sentence is that the former does not have a complementary that could occur at the end of the sentence[Chart]. See Figure.3.

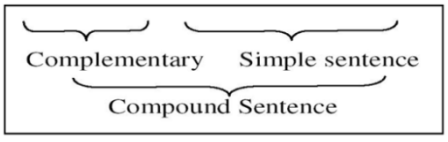


Figure 3. Simple and compound sentences

In Figure.4 a brief example of the syntactic structures of a nominal sentence (Figure.4.a) and a verbal sentence (Figure.4.b) according to the developed grammar are

given[Formal] .In Arabic, the definite article is a prefix, there is no copulative verb and direction of writing is from right to left[Formal] .

Here we are concerned with the nominal sentences because – somehow – it's synonymous for NPs.

In[6] the grammarian illustrates that the nominal sentence consists of two parts subject (Theme) (labeled `mubtadaa` in Arabic) and complement (Rheme) (labeled `khabar` in Arabic), each one of them has many cases :

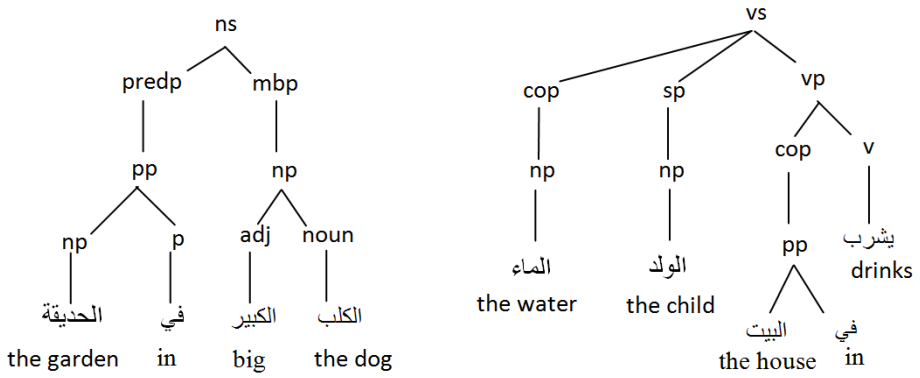
- **Theme:** can be:

Proper Noun (اسم صريح), Pronoun (ضمير), Demonstrative Pronouns (اسم إشارة), Relative Pronoun (اسم موصول), Interrogative Pronouns (اسم استفهام), Conditional Particle (اسم شرط) …

- **Rheme:** also has many cases:

1. Single (اسم مفرد): a noun for example.
2. Sentence (جملة): it could be nominal or verbal (contains a verb then subject then an object- optional-) sentence.
3. Phrase (شبه جملة): it could be prepositional (جار ومجرور) or adverbial (ظرفية) phrase.

Arab grammarians have not clearly determined what specific purposes or intentions might prompt the language user to choose a `mubtadaa` (theme) despite the presence of a verb (as illustrated in part 2 from the Rheme). It can be argued that whether the subject precedes or follows its verb is a question of communication rather than structure i.e. the reason for fronting a subject must be sought in the communicator's intention to concentrate attention on a noun phrase[16].



a) An example of nominal sentence (b) An example of verbal sentence

Figure 4. Examples on Arabic Sentences

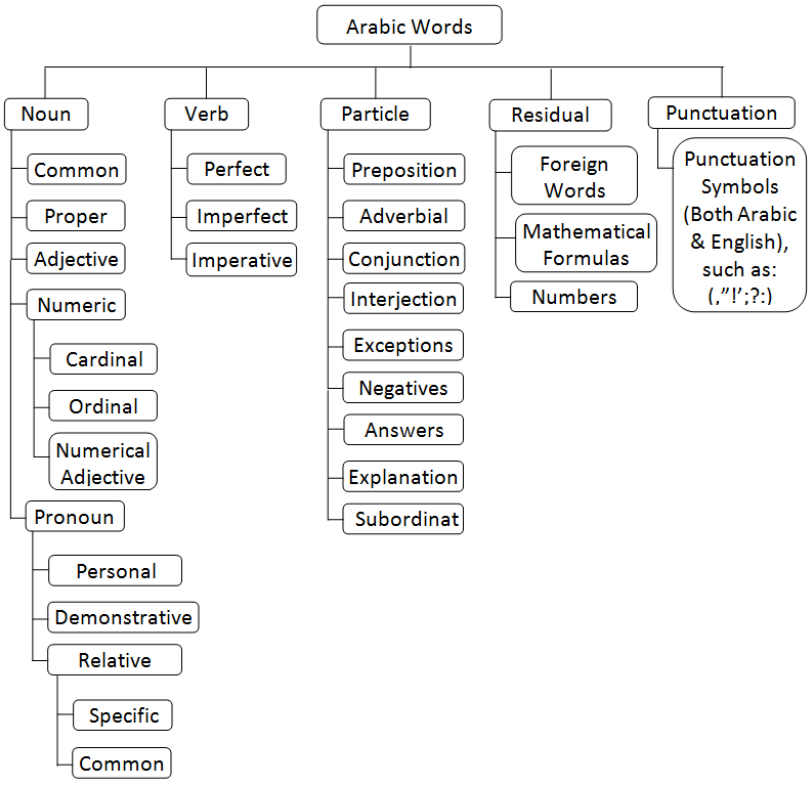


Figure 5. Arabic Part of Speech Categorization

4. Methodology

This section discusses the architecture of the Arabic Noun Phrase Extractor (ANPE) that we proposed, including exhibition for the steps of the design and implementation of the system. As mentioned in section 2 that most of the Noun Phrase extractors use three main steps:

- Tokenization
- Tagging
- Forming Noun Phrases

Our system also followed these steps in its design – as seen in Figure.6 –, so that they will be explained in details in the following sections.

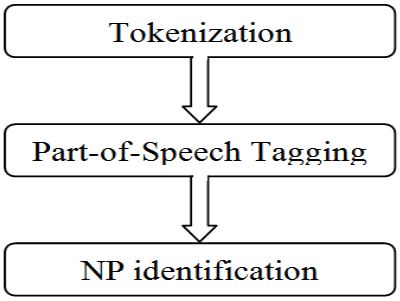


Figure 6. General steps of the NPE system

Here we proposed a phrase extracting system. In this section we described the effect of using such technique in the Information Retrieval System and introduced the results and evaluation of the system.

4.1. Building the IR System

The following steps have been used to implement the IR System:

Step.1: Remove all of the stop words from the documents.
Step.2: Build the inverted file for the documents.
Step.3: Choose a query (target query) from the source query list.
Step 4: Begin the search for the relevant documents to the selected query, the search will be done on the inverted file (for more details read the part talking about the inverted file).
Step 5: Use the cosine similarity formula (given below) to determine the similarity between the Query and the retrieved documents.

4.2. Evaluation of IR System

To evaluate the performance of the system built in 6.1, we will use two popular retrieval evaluation measures: Recall & Precision.

The Recall is calculated according to the following formula:

$$Recall = \frac{No.\ of\ retrieved\ relevant\ documents}{No.\ of\ relevant\ documents}$$

The Precision is calculated according to the following formula:

$$Precision = \frac{No.\ of\ retrieved\ relevant\ documents}{No.\ of\ retrieved\ documents}$$

Here the two measures will be applied on both systems we obtained, the system which uses the single terms as an indexing units (built in 6.1), and the system which uses the phrases (built in Section 4).

5. Conclusions

The discussion of current development approaches of Arabic noun phrase extractors illustrates the importance of the concept of roots and its rich structure that can be computerized. We have described the initial step of an ongoing research effort towards building Arabic Noun Phrase Extractor system, the next phases of this research are in progress, and we used to build it three common steps: tokenization; part-of-speech tagging; noun phrase identification.

REFERENCES

[1] "Modern Information Retrieval", R. Baeza-yates and B. Ribeiro-Neto, second edition, 2011

[2] Gerard Salton: Syntactic Approaches to Automatic Book Indexing. ACL 1988: 204-210

[3] Gerald Kowalski, Information Retrieval System—Theory and Implementation, Kluwer Academic Publishers (1997)

[4] Djoerd Hiemstra, "Information Retrieval Models'', In: Ayse Goker and John Davies (eds.), Information Retrieval: Searching in the 21st Century, *Wiley*, 2009.

[5] "NLP & IR a tutorial presented at EACL", Alan F. Smeaton, Dublin City University, 1995

[6] "اللغة العربية"، د.اسامة الخامش، النحو (الجزء الثاني) http://www.angelfire.com/tx4/lisan/grammar22.htm

[7] "Extracting Noun Phrases for all of MEDLINE", Nuala A. Bennett, Qin He, Kevin Powell, Bruce R. Schatz, 1999

[8] "Concept Extraction in the Interspace Prototype", Nuala A. Bennett, Qin He, Conrad T. K. Chang, Bruce R. Schatz, 1999

[9] "A tagset for the morph syntactic tagging of Arabic", Shereen Khoja, Roger Garside and Gerry Knowles, Lancaster University, 2001

[10] "Multilingual finite-state noun phrase extraction", Anne Schiller

[11] "A Phrase structure grammar of the Arabic Language", Ayman Elnaggar, 1990

[12] "APT: Arabic Part-of-speech Tagger", Shereen Khoja, 2001

[13] "A Chart Parser for Analyzing Modern Standard Arabic Sentence", Eman Othman, Khaled Shaalan, Ahmed Rafea, Cairo Univ., 2003

[14] "Arabic Language" http://www.student.virginia.edu/%7Earabweb/language.html

[15] "PhraseX and the SPECIALIST Minimal Commitment Parser", http://ii.nlm.nih.gov/MTI/phrasex.shtml

[16] "Language Journal of the Linguistic, Society of America",

1982

[17] "Acquisition System for Arabic Noun Morphology", Saleem Abuleil, Khalid Alsamara, Martha Evens, 2002

[18] "Coping with ambiguity and unknown words through probabilistic models", Weischedel, Ralph, Marie Meteer, Richard Schwartz, Lance Ramshaw, and Jeff Palmucci. (1993).

[19] "French Noun Phrase Indexing and mining for an Information Retrieval System", Hatem Haddad, France, 2000

[20] Fagan, Joel L. 1987. Experiments in Automated Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods. Ph.D. Thesis, Department of Computer Science, CorneU University.

[21] "Arabic Information Retrieval Perspectives", Ahmed Abdelali, Jim Cowie and Hamdy S. Soliman, 2004

[22] "Empirical Studies in Strategies for Arabic Retrieval", Jinxi Xu Alexander Fraser, and Ralph Weischedel, 2002