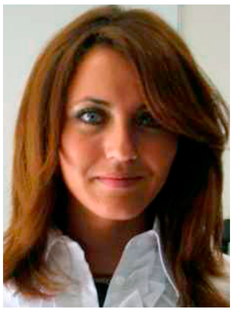
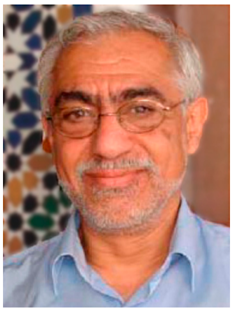
Improving Information Retrieval in Arabic through a Multi-agent Approach and a Rich Lexical Resource[†]

Mouna Anizi* and Joseph Dichy**

Université Lumière-Lyon 2 & ICAR Lab (UMR 5191-CNRS/Lyon 2), École Nationale Supérieure des Sciences de l'Information et des Bibliothèques (Enssib), Lyon, France,
*<Mouna.Anizi@univ-lyon2.fr>, **<Joseph.Dichy@univ-lyon2.fr>



Mouna Anizi is a PhD student at the Lyon 2-Lumière University and a member of the ICAR research Lab (CNRS/Lyon 2, ENS-Lyon). Her research interests include Arabic natural language processing, information retrieval, knowledge organization, and multi agent systems



Joseph Dichy is Professor of Arabic Linguistics at the Lyon 2-Lumière University and a member of the ICAR research Lab (CNRS/Lyon 2, ENS-Lyon). He is the author of a reference thesis on the writing system of Arabic (Lyon 2, 1990), of many works on Arabic descriptive and computational linguistics, on Medieval Arabic rhetoric and argumentation, and he is a recognized expert in the teaching of Arabic to speakers of other languages (TASOL). His main achievement in Arabic computational linguistics is his crucial contribution to the DIINAR lexical Arabic database.

Anizi, Mouna and Dichy, Joseph. **Improving Information Retrieval in Arabic through a Multi-agent Approach and a Rich Lexical Resource.** *Knowledge Organization, 38(5)*, 405-413. 18 references.

ABSTRACT: This paper addresses the optimization of information retrieval in Arabic. The results derived from the expanding development of sites in Arabic are often spectacular. Nevertheless, several observations indicate that the responses remain disappointing, particularly upon comparing users' requests and quality of responses. One of the problems encountered by users is the loss of time when navigating between different URLs to find adequate responses. This, in many cases, is due to the absence of forms morphologically related to the research keyword. Such problems can be approached through a morphological analyzer drawing on the DIINAR.1 morpho-lexical resource. A second problem concerns the formulation of the query, which may prove ambiguous, as in everyday language. We then focus on contextual disambiguation based on a rich lexical resource that includes collocations and set expressions. The overall scheme of such a resource will only be hinted at here. Our approach leads to the elaboration of a multi-agent system, motivated by a need to solve problems encountered when using conventional methods of analysis, and to improve the results of queries thanks to a better collaboration between different levels of analysis. We suggest resorting to four agents: morphological, morpho-lexical, contextualization, and an interface agent. These agents 'negotiate' and 'cooperate' throughout the analysis process, starting from the submission of the initial query, and going on until an adequate query is obtained.

Received 31 May 2011; revised 16 June 2011; accepted 20 June 2011.

[†] The authors would like to express special thanks to Dr. Mohamed Hassoun, professor of Information Sciences at the École Nationale Superieure des Sciences de l'Information et des Bibliothèques, who co-heads the SILAT resarch team together with Joseph Dichy (professor of Arabic Linguistics at the Lyon 2-Lumière University). The present work owes him a lot.

1.0 Introduction

1.1 Background: computer engineering and language

Natural Language Processing (NLP), currently oscillates, even today, between approaches related to engineering (computer sciences, statistics...), and approaches involving both linguistics and computer science (computational linguistics). Information retrieval, which is the central topic of the present paper, offers a particularly interesting example of collaboration between these two approaches. Let us consider, by way of comparison, automatic translation, which is also one of the most important areas of NLP. The last few decades have witnessed the development of systems based on rules and lexical resources, to which optimization methods based on statistics (frequencies), have later been added. The best known example is SYSTRAN, a leading supplier of language translation software (www.systran.fr). Over the past decade pure statistical systems have appeared, built on learning procedures through automatic matching from bilingual parallel corpora. REVERSO, a free online translation dictionary offers a familiar example (www.reverso.net). Another well-know example is the Google translator. Both systems provide users with Arabic-English online translation.

Statistics-based systems, after a remarkable phase of success, encounter two general types of difficulties:

- Available translators do not seem to be either complete or perfectible. Extending the corpora from which they have been constructed leads to a quantitative increase of data, but not a significant improvement in the quality of translations. Everything happens, so to speak, as extending corpora increased the quantity of data while the percentages of the structures statistically extracted from them remained constant.
- These systems are inadequate in specialized translation, unless the basis of which the corpora are built of includes a significant amount of specialized texts. To include new specific areas, one needs new corpora which prove to be expensive to acquire and preprocess. The economic benefit of this approach, compared to that based on rules and lexica, plus statistical optimization, is thus significantly weakened.

1.2 Specific introduction: the object of research

In this paper, we emphasize the need for a balanced collaboration between linguists and computer scientists in the field of information retrieval in Arabic. Such collaboration has been part of the work of the SILAT (Systèmes d'information, Ingénierie Linguistque et Traduction http://silat.univ-lyon2.fr) research team from the outset. During the 1990s this collaboration has allowed, jointly with the Tunisian research center IRSIT (Institut Régional des Sciences de l'Informatique et des Telecommunications), the achievement of the DIINAR.1 database (Dictionnaire Informatisé de l'Arabe, version 1 http://diinar.univ-lyon2.fr) (Dichy et al. 2002; Dichy and Hassoun 2005).

In the context of information retrieval in Arabic, keyword applications can prove to give poor results, which often, in addition, include noisy or ambiguous answers. Difficulties can be traced back to variation on several levels, including:

- Morphological changes, as in قلم, *qalam*, "pen" (sing.) ↔ أقلام , *'aqlâm*, "pens" (plural).
- Lexical variation, when two different words share a similar meaning (with shades of a difference), e.g.: بنت , *bint* and فتاة , *fatât*, "girl."
- Semantic variations due to homography, for example: عملة, respectively *'amala&*, "employees," "workers" or *'umla&*, "currency."

The writing system of Arabic features a very high level of homographic ambiguities (Dichy 1990). All of these variations have been included in the DIINAR.1 database.

To tackle such difficulties, we use a multi-agent approach. The approach consists of building models based on distributed artificial intelligence (DAI), simulating the "collaborative work" between human experts or cognitive modules implemented by a single expert. The use of language resources such as lexico-contextual databases in which lexical entries are associated with their contexts and expected (or 'preferred') collocations is a crucial aspect of our approach.

Section 2.0 below discusses information retrieval in Arabic, and highlights the need to develop a new lexical resource rich enough in collocations to solve a significant set of ambiguities. Section 3.0 recalls the multi-agent approach and its contribution to NLP, then describes the agents of our system and the way in which they cooperate to help users with the formulation of their queries.

2.0 The problem of information retrieval in Arabic

Written Arabic is highly ambiguous. The writing system of the language graphically realizes as mere diacritics vowels: consonants repetition, a great part of case-endings markers in nouns and of moods in verbs, etc. These signs are omitted in standard texts (correspondence, newspapers, literature, essays, scientific works, administrative documents). Information retrieval based on single word queries will consequently include a higher percentage of ambiguous queries than in English or French.

To deal with such problems, word level analysis needs to be based on a linguistic model. In the morpho-lexical theory (MLT) underlying the DIINAR.1 resource and the related analyzers, such a model includes a morpho-lexical resource. Entries are associated with word-level morpho-syntactic specifiers 'managing' relations between the lexical entries, which come within the word-form, in stem position (Dichy 1997), and the other word formatives. As a consequence of the complexity of the linguistic system, many queries based on a given Arabic word in standard nondiacriticized (or 'unvowelled') script unavoidably relate to several different words.

Such relations may, in addition, be both morphographic and semantic. For instance, the unvowelled word-form *Swt* صوت may refer to:

- *Sawt* "voice" (in both meanings: in Arabic and English [as well as in French], this word may correspond either to a 'sound' or to a 'vote');
- *Sawwata* "make a noise or a sound," "vote" (unlike the previous reading, there is no single word translation to English or French for the first meaning);
- *Suwwita*, passive of *Sawwata*.

This relatively simple example shows that, in applications based on the context-free processing of single word-forms, ambiguity might be due to:

- various diacriticizations supported by a great proportion of single written words (e.g. *Sawt / Sawwata / Suwwita*); or
- polysemy (two meanings in English and Arabic for *Sawt*, i.e. "voice" and "vote").

2.1 The lack of morphologically related forms and the reason

Let us consider four levels of automated analysis, traditionally corresponding to the first four 'layers' of a text: morphological (i.e. word level), lexical, syntactic (i.e. sentence or phrase level), and semantic, analysis. At what levels of indexing do current search engines operate? When one types a given keyword, the engine searches its database for all the Web pages that contain this word: no intelligence in the process, but a simple recognition of strings, which must be identical. In some cases, the engine appears to remove clitics, such as articles, prepositions, etc. Currently, most search engines (such as Google, Alta Vista, etc.) are still fundamentally based on the first level above, i.e. on words. Most search engines do not include tools devised for displaying morphologically related word-forms. Indeed, when a user initiates a query, then responses only match the query word, most often in the form in which it has been entered, or in a very close form, such as the word with the article al-. Thus, a query on Google (French version) launched in December 2010 for the conjugated verb form *partirais* ("would leave") gave approximately 285,000 pages in French, obtained in 0.14 seconds. The only variation that we found is *partirai* ("will leave") (without the "s" of the conditional).

The lack of links between a given word and morphologically related forms (e.g.: in French: the participles *partant* and *parti*, the infinitive *partir*, the noun *départ*, "departure" in English, as well as the conjugated forms of the verb) must be related to what is widely known as Google page-ranking, which is based on extensive exploration of connections in the network (crawling) and much progressive indexing, which goes up from one month to another with remarkable speed (a synthesis has been presented in Anizi and Dichy 2009; see also http://www.rankspirit.com and Peyronnet 2007).

The aformentioned result, for *partirais* ("would leave"), is impressive, considering the speed with which Google results are displayed. Let us suppose the engine had been coupled with a morphological analyzer. Whatever the speed of the analyzer, and given the hundreds of millions of words that need to be analyzed (or even more, since test analysis must be compared to the query), this would have resulted in a very significant, if not disastrous, slowdown. It is absolutely out of the question, for a search engine, to work for many minutes or even for hours: for example, by granting morphological analysis, on a purely theoretic basis, 1/10,000th of a second, the approximately 40 million responses we obtained almost instantaneously in April 2009 for the query *kawkab* "aster, celestial object," would have required waiting for 4,000 seconds, i.e. for about 1:06 hour. In addition, integrating morphological analysis in exploration operations (crawling) and automatic indexing would result in considerable slowdowns entailing a heavy disruption of operations related to Google ranking. Consequently, we do not, in any event, consider integrating an analyzer in Google searches, but rather propose a method for the enrichment and the reformulation of queries.

Morphological variations associated with the query keyword can, in many cases, be very useful to users. A person seeking in Arabic the word *safîr*, "ambassador" (9.9 million Google results in 0.23 seconds on November 18, 2010), is likely to need results related to the feminine form *safira* or the 'broken plural' form *sufarâ'*. A recent search on Google for the same word has nevertheless given only one form, the masculine singular, with or without the article al-.

The result is the same for the compound word *'alim (al)-'âtâr'*, "archeologist": a query launched in November 2010 gave approximately 4.46 million results in 0.22 seconds, which did not include *'âlimat (al)-'âtar'* ("archeologist, fem. ") or *'ulamâ' (al)-'âtar* (plural masc.). For *qânûn al-jinâya*, "Criminal Law" (about 291,000 results in 0.25 seconds on November 19, 2010), we only got the expression determined by the article al- (before *jinâya*), in the singular form. In other words, the only answers obtained were those related to the original form of the query. No results were given for the plural *qawânîn* or the indeterminate form *qânûn jinâya* (without the article al-). To obtain the latter, a given user must proceed with new queries in the plural form, without the article, etc.; i.e., build in his or her own mind the set of morphologically related word-forms potentially needed.

2.2 Problems due to contextual and semantic variation

A query on Google with the word *al-mal*, "money" produced 9,690,000 results in September 2008 and approximately 52.6 million ones in November 2010, which illustrates both the increase in the number of users of Arabic language and the subsequent development of the exploring (crawling) and automatic indexing processes of Google page-ranking. In neither query, though, did the plural *'amwâl* appear. Note that one can observe the same lack of formal relationship between the French word *capital* and its plural *capitaux*, which can be assigned to the same reasons as those given above for Arabic.

Here are some examples of results for the word al-mal:

– *al -'azma l-mâliyya* "the economic crisis;"
– *al-'awrâq al-mâliyya*, word for word: "the commercial papers" (papers related to money), "paper money," "banknotes;"
– *al-mu'assasât al-mâliyya*, "financial institutions;" and,
– *al-'aswâq al-mâliyya / al-sûq al-mâliyya*, "financial markets" / "the financial market."

Noticeably, the suffix *–iyy* of the relative adjective or noun, appears here with the noun *al-mâl*, in the results of the query, together with the article, unlike other word-form variations.

The last example is interesting to observe. With the Google request related to *al-mal*, we get two answers *al-'aswâq al-mâliyya* and *al-sûq al-mâliyya*, with the singular and plural forms of *sûq*, "market." However, by bringing the application on either *sûq* (sing.) or *'aswâq* (plural), we only obtained one answer, respectively *sûq al-mal* ("the financial market"–word-for-word: "market of the money" or "currency") or *al-'aswâq al-mâliyya* ("the financial markets"). Albeit both expressions appear in the same context, Google treats them as two independent results. This highlights the need for collaboration between morphology on the one hand and contextualization on the other.

The query for the word *safîr* is interesting for another reason. The word means, according to the context, "ambassador" (literally) or "symbolic representative" (in a figurative sense, very common in Arab newspaper writing today). Examples:

– *al-safîr al-jazâ'iriyy*, "the ambassador of Algeria" (literal meaning);
– *safîr al-nawâya al-Hasana*, "the ambassador of good intentions" (figurative meaning).

Also found with the figurative meaning:

– *muntada safîr al-Hubb*, "the [internet] forum 'the ambassador of love';"
– *safîr al-Hubb*, "the ambassador of love" (name of a website, www.sfiiral7b.com) – compare with the similar English re-use of the word ambassador);
– *muntadayât safir al-waTan li-l-jamaahiir al-ahlawiyya*, forum of supporters of the Egyptian football club al-Ahli, word-for-word "Forum of the country's ambassador to the public Ahliotes;" and,
– *safir al-shawq*, "the ambassador of nostalgia."

These examples show that for a given query, users obtain results that could be distinguished, in order to eliminate unwanted answers, using a contextualization tool, based on a lexico-contextual resource that includes collocations and set expressions. The question is: how?

2.3 Developing a new type of lexical resource

To achieve this, the development of a new lexical resource is needed. This resource will be based initially on DIINAR.1 (computerized dictionary of Arabic version 1), whose first foundations were laid in Hassoun (1987), and Dichy (1990). The extension prototype we plan to build will be based on a linguistic model consisting of a set of lexical information, the types and hierarchy of which will be defined in a DTD (Document Type Definition) corresponding to a valid XML file based on the FLEXARABE ('format lexical de l'arabe') lexical format of Arabic elaborated by Dichy (1990). (A first prototype based on the same work was introduced in Anizi 2008).

In an experiment conducted with the Google search engine in November 2010, the Arabic word *kawkab* gave about 56.4 million answers in 0.16 seconds. This word has several senses: "planet," "star," "movie star," etc. Regardless the user may only need one of these meanings.

Collocations such as *kawkab al-marrîkh*, "the planet Mars," *kawkab al-shams*, "the Sun" or *kawkab al-sharq*, "the Star of the Orient" (traditionally referring to the great singer Umm Kulthum), etc., correspond to specific contexts that can be stored in a database and provide users with a pertinent response.

The software we propose will guide users in their searches on Google, Yahoo!, Bing, Ask, AltaVista or any other search engine that includes Arabic searches. The device should offer users functions allowing the morphological optimization of queries and their semantic filtering. To build our lexico-contextual resource, we will use representation tools belonging to the XML/OWL galaxy as well as processing tools (scripting languages such as Perl or XSLT) applied to corpus treatments.

Let us now turn to the outline of our approach.

3.0 Morphological And Lexical Analysis, As Part Of A Multi-Agent System

Let us first recall in few words what a multi-agent system (MAS) is. A MAS can be described as a set of interacting autonomous agents (Erceau and Ferber 1991; Ferber and Gasser 1991; Gleizes and Glize 1990); an agent is a real entity (robot) or an abstract one (a software module) located in an environment in which it is able to act. This entity has a capacity of perception and a partial representation of its environment. It can 'communicate' with other agents. Its autonomous 'behavior' is a consequence of its 'observations,' stored or acquired knowledge, and interactions with other agents. An agent is, thus, an entity that is capable of acting 'rationally' and of 'intentionally' meeting its own 'goals,' and reflecting the current state of knowledge.

MAS systems differ according to criteria such as: the type of agents (reactive vs. cognitive), the type of agent behavior (selfish vs. altruistic; cooperative vs. confrontational), the communication mode (communication through shared memory, communication by messages, communication through environment), type of control (centralized vs. decentralized control) (Warren 1998) and the settlement structure (the two main types of architecture are the 'blackboard' and the architectures based on agent languages).

The distributed resolution of a problem depends on several factors including:

- the type of agents used (cognitive, reactive);
- the type of control implemented (centralized, decentralized or mixed);
- the type of agent behavior (cooperative vs. confrontational);
- the technical implementation of interactions that can be managed by data;
- the sharing of results led by goals; and,
- task-sharing, or mixed interactions.

Resolution by sharing tasks runs as follows: decomposition of the problem into subtasks, distribution of subtasks, resolution and integration of results. Resolution by sharing results requires a distribution of knowledge, and a summary of results (Warren and Stefanini 1996). The system we propose uses a functional approach, 'cognitive' and 'altruistic' agents and a decentralized control. In case of conflicts, a process of 'negotiation' between agents can be launched. Resolution in our system uses a basic technique for the implementation of interactions between agents which is the intentional communication by sending messages (3.2 below).

3.1 Contribution of the multi-agent approach to NLP

Natural language processing (NLP) is often seen as a set of disjoint and successive processes, rarely as calculations that can be performed in parallel, and much less often as knowledge and collaborative processes.

Classical NLP methods, which propose going through a morphological or morpho-lexical stage (word level), a syntactic stage (sentence level, considered from a formal point of view), and a semantics stage, have helped solve several problems and develop useful systems in many applications. Well known examples are those of spelling and grammar checking. One can also cite the progress of automatic translation and computer-assisted translation. Nevertheless, these methods encounter limitations in terms of matching linguistic analysis levels, lack of interaction between representation levels, lack of distribution of control and knowledge, and difficulties in modifying their system. The main consequence of working with disjoint levels of analysis is the risk of combinatorial explosion.

Distributed architectures allow improving results through a better collaboration between different levels of analysis. Thus the boundaries are less clear-cut than in a sequential and modular system, where each level is isolated from the others. One of the contributions of the MAS approach is cooperation, which aims at activating the proper analysis of a text by eliminating parasite solutions in order to achieve a robust and optimal analysis (Warren and Stefanini 1996; Aloulou et al. 2002). Robustness is frequently sought in morphological and syntactic analyzers. In Arabic, 'well prepared' texts for the sake of NLP applications remain rare. Almost no text, for instance, currently includes consonant doubling marks (*shadda*), the lack hereof is responsible for a large percentage of NLP analysis difficulties (Abbès and Dichy 2008). We believe Distributed Artificial Intelligence (DAI) and in particular, MAS approaches, can contribute powerfully to the solving of such general or language-specific NLP problems.

3.2 Collaboration between the morphological, morpho-lexical and contextualization agents

Agents in our approach are defined according to the tasks assigned to our system, which are in turn described according to an analysis of human behavior in performing similar tasks. Among the agents that are included in our modeling, we consider the morphological agent, the morpho-lexical agent, the contextualization agent, and the user-friendly interface agent.

3.2.1 The morphological agent

This agent segments and analyzes word-forms, context free (Dichy and Hassoun, eds, 1989). For instance:

- al-mâlu, "the money," after segmentation breaks down to the following formatives: *al-*, article, *mâl*, stem, "money,"*–u*, case-ending suffix, nominative;
- *wa-mâl-a-hu*, "and his money," is analyzed as: *wa-*, coordinating conjunction, *mâl*, stem, "money,"*–a*, case-ending suffix, accusative, *–hu*, enclitic complement pronoun, masc. sing. "of him."
- *mâl-iyy*, "financial," gives: *mâl*, stem, *–iyy*, relative adjective suffix.

Such relations can only be accounted for if the morphological analyzer is based on a lexical database enriched with morpho-sytactic specifiers. These specifiers are the core of the morpho-lexical theory introduced in (Dichy 1990, chap. X) and (Dichy 1997). For instance, the relative adjective suffix is not compatible with all nouns, e.g.: *kitâb*, "book," has no relative adjective (meaning "book-like" or "bookish"), because the form *kitâbiyy* ("scriptural," "in-writing") is related to *kitâba&*, ("writing").

There is a need, in addition, for a singular ↔ 'broken plural' relation, in applications such as information retrieval and automated translation, e.g., *mâl* (singular) ↔ *'amwâl* ('broken plural').

The morphological agent is designed to determine the morpho-syntactic characteristics of each word using the DIINAR.1 resource, in which each entry is associated with a set of specifications called W-specifiers, i.e. specifiers operating at word-form levels (Dichy 1997).

We must also consider two types of association between the core lexical formative (or stem) of the word-form, and other word formatives, such as the suffix *+a&*. For instance, the word-form *jâmi'a&* supports two analyses:

- the active participle *jami'*, "bringing together," followed by the feminine suffix *+a&*;
- the noun *jâmi'a*, "university," in which the suffix *+a&* is not a femine ending, but a lexicalized extension-formative, which cannot be removed, and actually belongs to the lexical unit (see, for this analysis, (Dichy 1997)).

Such distinctions are crucial in the contexts of automated translation or information retrieval. They can only be found in a morpho-lexical database. Collaboration within this agent between the DIINAR.1 database and the rules implemented in the analyzer yields all the relevant word-level information corresponding to a given word-form placed at the input of the analyzer.

3.2.2 The morpho-lexical agent

This agent must work with a grammar and a lexicon. It determines whether a word does or does not belong to the lexicon and aims at associating each form with one or more lexical inputs and one or more lexical categories (with the values of their variables). The morpho-lexical agent also relates the word with its other morphologically related forms, eg. *sakana* (perfective, "he dwelt") ↔ *ya-skun-u* (imperfective, "he dwells") ↔ *sakanun* (infinitive form, 'dwelling'), through consulting the DIINAR.1 resource. The plural form can also be used to identify the meaning, e.g., *'âmil* (عامل) has two plural forms, *'ummâl* (عمال): 'workers, laborers' and *'awâmil* (عوامل) 'factors.'

3.3.3 The contextualization agent

This agent needs the knowledge of both the morphological and morpho-lexical agents, on the basis of which it reads the new XML-based vocabulary that includes the usual contexts related to a lexical unit, and connects the word with the set expressions, idioms, collocations, "preferred contexts," etc. in which it is included. This agent aims at disambiguating words according to their meaning, through helping users in choosing the exact context that matches their need.

The contextualization agent is based on a typology of syntactic contexts, such as 'annective structures' (in Arabic, *mudâf* and *mudâf 'ilayhi*), the 'name/adjective sequence' (*na't wa -man'ût*). The example above of the word *'âmil* shows the existence of two different meanings, related to morphological variation (here: sing. ↔ plur. relations). These meanings are also related to specific contexts, e.g. *'amil binâ'*, "construction worker" (plural *'ummâl binâ'*) vs. *'amil 'iqtiSâdiyy* "economic factor" (plural *'awâmil 'iqtiSâdiyya*).

To disambiguate the query, users can choose the appropriate context with the help of the interface below. In addition, the example features in the plural, converging results between the morpho-lexical and contextualization agents. The prototype of a new lexico-contextual database rich in contexts was built in XML, on the basis of Dichy's (1990) lexical format.

3.2.4 The user-friendly interface agent

This agent is not parallel to the other three, which are designed to 'cooperate'. It provides users with the benefits of an interactive system assisting them in the reformulation of queries. It is not based, as are the preceding agents, on analyzers drawing on lexical resources, but rather, on the observation of the behavior of actual users.

The system plays an important role in suggesting words and forms, and displaying on computer screens lists of words and expressions. The user examines the list and decides on the choice of lexical units that he or she wishes to add to the query. The final decision in the selection thus belongs to users.

Let us now consider the different phases of the agents based on morphological, lexical and contextual analysis, and the way in which they interact.

3.3 Communication between agents

In the course of segmenting the query word into formatives, the morphological agent asks the morpho-lexical agent about its various virtual context-free segmentations. Thanks to the lexical knowledge extracted through consulting the DIINAR database, the morpho-lexical agent returns a set of linguistic information such as the grammatical category of the stem included in the word-form, its gender, number, etc., and other features associated with either the stem itself, or its prefixes and suffixes. The morphological agent then retains existing words, and rejects invalid segments or parasite analysis (i.e. word-form analysis that would have been proposed by the morphological agent, for forms not evidenced by the language, stored in the DIINAR.1 database, which allow excluding non-attested 'virtual' forms). The morpho-lexical and morphological agents send the analysis retained to the contextualization agent, which in turn associates each of them with the set of its collocations, through consulting the new lexical resource. The contextualization, the morpho-lexical and the morphological agents work together, thanks to the technique of sharing results, to merge the results and enrich the user's request. Writing the new query thanks to the user-friendly interface finally helps users in formulating queries. We will aim to add to search engines a 'smart' tool for the generation of forms (see examples in Anizi and Dichy 2009).

4.0 Design And Modeling

We are in the course of developing, for the purpose of the multi-agent communication described above, models of the communication that can occur during the analysis of a given query (composed of one word or more). These models require a detailed design of our MAS system. For this we use AUML (Agent

Unified Modeling Language) which is an extension of the UML method. The latter is mainly used for modeling multi-agent interactions, and mainly introduces two new concepts in the representation of protocols and representation levels of the interactions between agents (Odell et al. 2000).

Our model has not yet been entirely defined (Figure 1. below presents an outline). However, we retained the option of a functional approach of modeling. This makes it possible to distribute knowledge between agents specialized in their respective areas, called expert agents.

5.0 Conclusion

The first part of this paper presents problems inherent to the Arabic language in relation with both word-form analysis, and information retrieval. The second part proposes the modeling of a tool for information retrieval in Arabic that can help reformulating users' queries and supporting them in finding relevant information. The modeling is based on problems encountered in effective queries on the Google search engine. To overcome the problem of morphological changes our interface exploits the basic forms of words in the query produced by the morphological analyzer to infer derived forms. To disambiguate and contextualize the query we elaborated the prototype of a lexical resource, which is rich in collocations, set expressions and contexts and aims at assisting users in formulating their queries.

We chose a multi-agent approach (parts three and four) to ensure collaboration between morphological, lexical and contextualization modules, by bringing together several expert agents. The crucial aspect of this model is the combination of inter-agent communication and the provision of rich lexical resources. Generally speaking, the system we propose, regarding its configuration, can be integrated in many Arabic language engineering applications such as passage retrieval in question/answering systems (Abouenour et al. 2010).

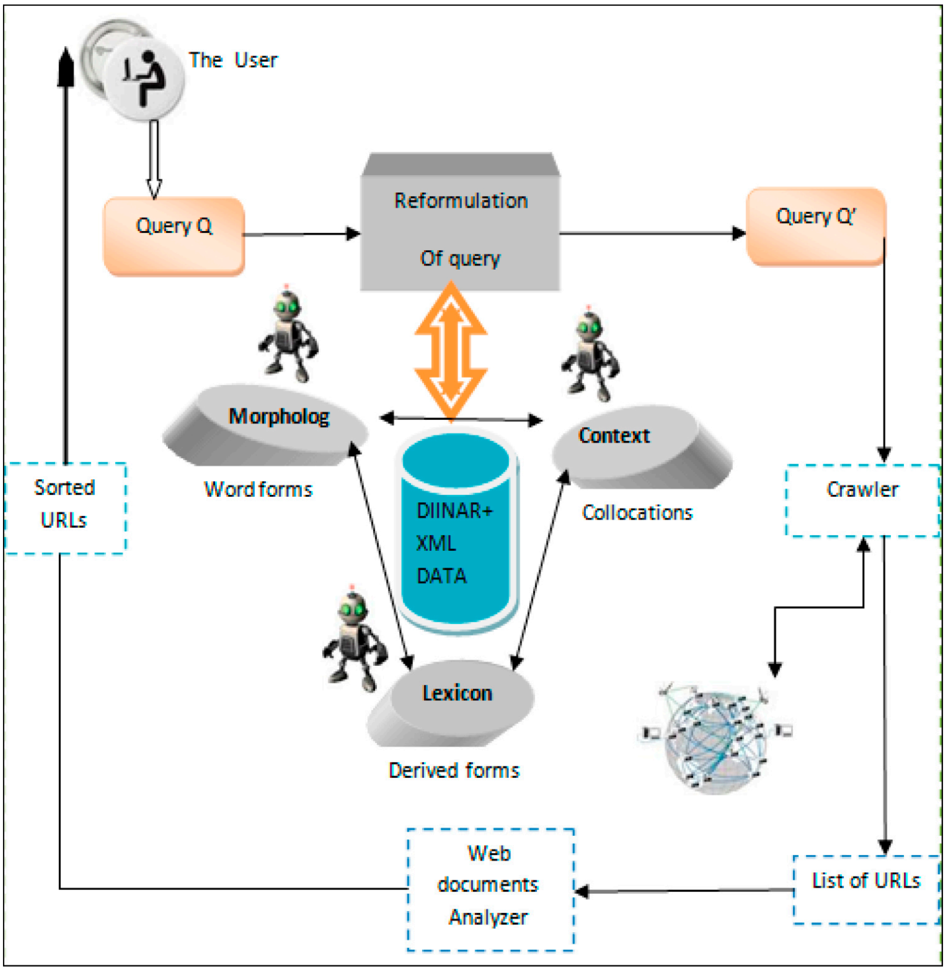


Figure 1. Overview of the system

References

Abbès, Ramzi and Dichy, Joseph. 2008. Extraction automatique de fréquences lexicales en arabe et analyse d'un corpus journalistique avec le logiciel AraConc et la base de connaissances DIINAR.1. In Heiden, Serge and Pincemin, Bénédicte *JADT 2008: proceedings of 9th International Conference on Textual Data statistical Analysis, Lyon, March 12-14, 2008*. Lyon: Presses Universitaires de Lyon, vol. 1, pp. 31-44. Available: http://www.cavi.univparis3.fr.

Abouenour, Lahsen; Bouzouba, Karim and Rosso, Paolo. 2010. An evaluated semantic query-expansion and structure-based approach for enhancing Arabic question/answering. *International journal on information and communication technologies* 3 no. 3: 37-51.

Aloulou, Chafik; Belguith, Lamia and Ben Hamadou, Abdelmajid. 2002. MASPAR: Multiagent System for Parsing Arabic, IEEE International Conference on Systems, Man and Cybernetics, vol. 7, pp. 6-9, Hammamet-Tunisie, Octobre 2002.

Anizi, Mouna and Dichy, Joseph. 2009. Assessing Word-form based Search for Information in Arabic: Towards a New Type of Lexical Resource. In Choukri, Khalid and Maegaard, Bente *Proceedings of the Second International Conference on Arabic Language Resources and Tools, 22-23 April 2009*. Cairo: The MEDAR Consortium. I can't find the proceedings themselves to get page #s.

Anizi, Mouna. 2008. *Structuration du lexique arabe en vue de faciliter la recherché d'information.* Mémoire de Master 2, Lyon 2.

Dichy, Joseph and Hassoun, Mohamed eds. 1989. *Simulation de modèles linguistiques et Enseignement Assisté par Ordinateur de l'arabe - Travaux SAMIA I*. Paris: Conseil International de la Langue Française.

Dichy, Joseph and Hassoun, Mohamed. 2005. The DIINAR.1-« معالي » Arabic Lexical Resource, an outline of contents and methodology. *The ELRA newsletter* 10n2: 5-10.

Dichy, Joseph, Braham, Abdelfattah, Ghazali, Salem and Hassoun Mohamed. 2002. La base de connaissances linguistiques DIINAR.1 (DIctionnaire INformatisé de l'Arabe, version 1). In Braham, Abdelfattah, ed. *Actes de la conférences internationale sur le Traitement automatique de l'arabe, Proceedings of the International Symposium on The Processing of Arabic, Tunis (La Manouba), 18-20 April 2002*. Tunis: Université de La Manouba, pp. 45-56.

Dichy, Joseph. 1990. *L'Écriture dans la représentation de la langue: la lettre et le mot en arabe*, thèse d'État, Université Lumière-Lyon 2.

Dichy, Joseph. 1997. Pour une lexicomatique de l'arabe: l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot. *Meta* 42: 291-306. Available: www.Erudit.www.erudit.org/revue/meta/1997/v42/n2/002564ar.pdf

Erceau, Jean and Ferber, Jacques. 1991. L'Intelligence Artificielle Distribuée. *La Recherche* 233: 750-58.

Ferber, Jacques and Gasser, Les. 1991. Intelligence artificielle distribuée. In (author or editor?) *Proceedings of the 11th International Workshop on Expert Systems and their Applications, "Tools, Techniques & Methods," Avignon, 27-31 May 1991.* Place: publisher, pp. x-x.

Gleizes, Marie-Pierre and Glize, Pierre. 1990. Les systèmes multi-experts. Technologie de pointe no. 38. Paris: Hermès.

Hassoun, Mohamed. 1987. *Conception d'un dictionnaire pour le traitement automatique de l'arabe dans différents contextes d'application,* thèse d'État, Université Lyon 1.

Odell, James, Parunak, H. Van Dyke and Bauer, Bernhard. 2000. Extending UML for Agents. In Wagner, Gerd, Lesperance, Yves and Yu, Eric, eds. *Proceedings of the Agent-Oriented Information Systems Workshop at the 17th National conference on Artificial Intelligence, July 30 – August 3, Austin, TX.* Menlo Park, CA: AAAI Press, pp. 3-17.

Peyronnet, Guillaume. 2007. Comment fonctionne un moteur de recherche comme Google? Pomms, Internet. 1 January 2007. Available: http://www.pomms.org/comment-fonctionne-un-moteur-de-recherche-comme-google--121.html.

Warren, Karine and Stefanini, Marie-Helene. 1996. Modélisation et validation de protocoles de communication dans l'architecture TALISMAN. In *Proceedings of the International Conference on Natural Language Processing and Industrial Applications (NLP+IA 96), Moncton, Canada, 4-6 June 1996.* (City and publisher?) pp. 270-276.

Warren, Karine. 1998. *Gestion de conflits dans une architecture multi-agents d'analyse automatique de textes,* thèse de doctorat en informatique et communication. Université Stendhal-Grenoble III.