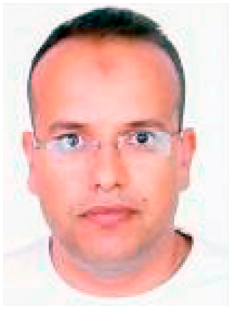
Organizing Contextual Knowledge for Arabic Text Disambiguation and Terminology Extraction †

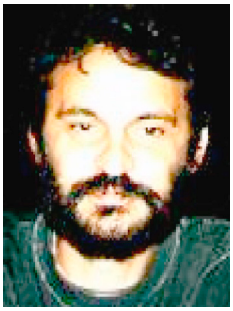
Ibrahim Bounhas*, Bilel Elayeb**,
Fabrice Evrard***, Yahya Slimani****

* Department of Computer Science, Faculty of Sciences of Tunis, University of Tunis, 1060 Tunis, Tunisia, <Bounhas.ibrahim@yahoo.fr>
** RIADI-GDL Research Laboratory, The National School of Computer Sciences (ENSI), 2010 Manouba, Tunisia, <Bilel.Elayeb@riadi.rnu.tn>, Informatics Research Institute of Toulouse (IRIT), 02 Rue Camichel, 31071 Toulouse, France.
*** Informatics Research Institute of Toulouse (IRIT), 02 Rue Camichel, 31071 Toulouse, France. <Fabrice.Evrard@enseeiht.fr>
**** Department of Computer Science, Faculty of Sciences of Tunis, University of Tunis, 1060 Tunis, Tunisia, <yahya.slimani@fst.rnu.tn>

Ibrahim Bounhas obtained a license degree in computer science applied to management in 2004 from the High School of Management of Tunis (ISG) and a master's degree in computer science in 2006 from the National School of Computer Sciences (ENSI). He is a PhD student at the Department of Computer Science of Faculty of Sciences of Tunis (2007-2008). He presented a master's thesis entitled "Un analyseur de contenu des documents scientifiques du Web." His current research interests are: ontology engineering, document analysis, and Arabic text processing.

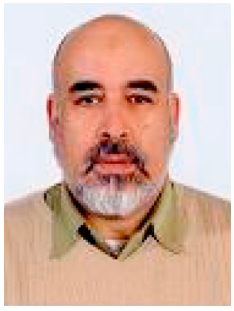


Bilel Elayeb is an assistant professor at the National School of Computer Science of La Manouba in Tunisia. He obtained his PhD in computer science from the National Polytechnic Institute of Toulouse and the National School of Computer Science, Tunisia in 2009. He obtained a master's thesis in computer science from the ENSI in 2004. His research focuses on information retrieval, computational linguistics, Arabic NLP, and multiagent systems, including possibility theory and hierarchical small-worlds networks. He has been a member of the RIADI research laboratory since 2002 and of the Informatic Research Institute of Toulouse (IRIT) since 2005.



Fabrice Evrard has been an Assistant Professor at ENSEEIHT, Toulouse, France since 1983. His research focuses on multiagent systems, dictionary modeling and analysis, information retrieval, computational linguistics, NLP, and hierarchical small-worlds neworks. He supervised a master's degree in artificial intelligence at the National Polytechnic Institute of Toulouse (INPT). He conducted Le Groupe Raisonnement, Action et Actes de Langage (GRAAL) team, which is a part of LILaC research group at the Informatic Research Institute of Toulouse (IRIT), France. He has supervised many master theses and PhD theses in artificial intelligence, information retrieval, and NLP.

Yahya Slimani studied at the Computer Science Institute of Alger's (Algeria) from 1968 to 1973. He received the B.Sc.(Eng.), Dr Eng and Ph.D degrees from the Computer Science Institute of Alger's (Algeria), University of Lille, and University of Oran (Algeria) in 1973, 1986, and 1993, respectively. He is currently a professor at the Department of Computer Science of Faculty of Sciences of Tunis. His research activities concern datamining, text mining, ontology engineering, parallelism, distributed



systems, and grid computing. Professor Slimani has published more than 90 papers from 1986 to 2008. He joined the editorial boards of the *Information International Journal* in 2000.

Bounhas, Ibrahim, Elayeb, Bilel, Evrard, Fabrice, and Slimani, Yahya. **Organizing Contextual Knowledge for Arabic Text Disambiguation and Terminology Extraction.** *Knowledge Organization, 38(6)*, 473-490. 38 references.

ABSTRACT: Ontologies have an important role in knowledge organization and information retrieval. Domain ontologies are composed of concepts represented by domain relevant terms. Existing approaches of ontology construction make use of statistical and linguistic information to extract domain relevant terms. The quality and the quantity of this information influence the accuracy of terminology extraction approaches and other steps in knowledge extraction and information retrieval. This paper proposes an approach for handling domain relevant terms from Arabic non-diacriticised semi-structured corpora. In input, the structure of documents is exploited to organize knowledge in a contextual graph, which is exploited to extract relevant terms. This network contains simple and compound nouns handled by a morphosyntactic shallow parser. The noun phrases are evaluated in terms of termhood and unithood by means of possibilistic measures. We apply a qualitative approach, which weighs terms according to their positions in the structure of the document. In output, the extracted knowledge is organized as network modeling dependencies between terms, which can be exploited to infer semantic relations. We test our approach on three specific domain corpora. The goal of this evaluation is to check if our model for organizing and exploiting contextual knowledge will improve the accuracy of extraction of simple and compound nouns. We also investigate the role of compound nouns in improving information retrieval results.

† Sincere thanks to Dr. Ryan Roth, Dr. Nizar Habash, Dr. Owen Rambow and all researchers from Columbia University, USA, who participated in developing MADA and helped us working with this tool. We would also like to thank the anonymous reviewers for their helpful comments and suggestions.

1.0 Introduction

The huge amount of knowledge present in documents needs to be organized to help the user exploit its richness. On the one hand, documents should be indexed to help search engines retrieve their content. On the other hand, there is a growing need for automatic text analysis, annotation techniques, and knowledge organizing systems (KOS) of several types (Bourigault and Lame 2002; Broughton et al. 2005). Any of these resources is structured as a set of units (terms or concepts) organized through various types of relations. Consequently, term extraction is an important step in Information Retrieval (IR) (Boulaknadel 2006), question answering (Ferret et al. 2002), knowledge extraction, and many Natural Language Processing (NLP) tasks. Candidate term extraction requires to define statistical measures to weight and to filter terms, but also to handle Multi-Word Terms (MWTs). According to Martínez-Santiago et al. (2002, 1), detecting these entities "can be successfully used in many different tasks." More precisely, the knowledge organization literature shows that noun phrases (NPs) are the best entities that represent the document's subject (Malaisé et al. 2003; Boulaknadel 2006). In this field, Souza and Raghavan (2006, 559) defend "the hypothesis that NPs carry the greater part of the semantics of a document." In addition, many ontology construction tools exploit networks of syntactic dependencies. In Bourigault and Lame (2002), a network of simple and compound noun phrases generated by a syntactic analyzer is enriched by distributional links to build a "documentary ontology" exploited as a thematic index to access documents.

Semi-structured documents (e.g., books, scientific papers, and encyclopedia) contain additional information which may be exploited to understand, to index, and to infer knowledge from corpora. This paper proposes to exploit such knowledge in terminology extraction. In fact, we transform the structure of documents, which represents a logical division of knowledge, into an empiricist contextual graph. Indeed, many researchers have investigated and continue to work on extracting candidate terms from textual and semi-structured corpora. However, only few works considered Arabic documents. This task requires sophisticated corpus analysis tools which are available for many languages (e.g., French and English). Despite the great work done in the field of Arabic NLP, existing ontology environments can not be directly used to process Arabic documents. One of

the main causes is the lack of sufficient linguistic resources for the Arabic language. Also, approaches for Arabic text disambiguation have to be improved since this language is highly ambiguous. Related work proves both the usefulness and the difficulty of building these resources (Attia et al. 2008). This difficulty made existing works adopt a manual approach (e.g., Elkateb et al. 2006; Zaidi and Laskri 2005; Attia et al. 2008) or a semi-automatic approach (Rodríguez et al. 2008). A great deal of work has been done in the field of Arabic text parsing (Attia et al. 2008) and morphologic disambiguation (Habash et al. 2009). These approaches perform only the first step required for term extraction. Consequently, they are unable to give a clear evaluation of candidate terms. Other works of interest to document indexing and term extraction lack sophisticated NLP tools (Larkey et al. 2002; Boulaknadel et al. 2008). Through this literature, we feel the need for an approach which exploits sufficient and well-organized linguistic and contextual knowledge to handle terms.

Probabilistic measures allow one to evaluate separately two fundamental properties of terms. For example, TF-IDF (Salton and McGill 1986) is used to evaluate termhood whether scores like LLR (Dunning 1994) are employed to compute unithood of compound terms. In this paper, we define a possibilistic measure for relevance which combines the termhood and unithood dimensions of terms.

When we consider non-diacriticised Arabic texts, this process generates many types of ambiguities. Morphosyntactic disambiguation and domain relevance evaluation were previously considered as two separated steps. Our possibilistic measure is used both for disambiguation and for domain relevance evaluation considered as interrelated tasks. Our approach exploits the structure of documents which constitutes rich contextual information. The document is seen as a tree where nodes are linked with structural relations. The relevance of a term which appears on a given node is related not only to its distribution in corpora, but also to the position of the node in the structure of the document and its structural relations. Because the context is composed of complex relationships, we model this problem as an IR task where the query encodes contextual constraints. These queries allow one to disambiguate syntactic trees and to retrieve the most domain relevant terms.

We test our hypotheses in the particular context of extracting many domain terminologies from books of Arabic stories organized by theme. Because of the lack of gold standards, the extracted terminologies are checked by human experts who build a reference list for each domain. This method is influenced by the subjectivity of the expert. That's why we suggest a second method of evaluation which consists of using the extracted knowledge in the context of a possibilistic IR system. We report encouraging results which are to confirm the targets set for the precision and recall metrics compared to the state-of-the-art measures.

This paper is structured as follows. Section 2 presents a literature review in the field of terminology extraction, focusing on the characteristics of the Arabic language. In Section 3, we present our approach for domain relevant term identification based on a critical study of existing approaches. We experiment this approach on Arabic corpora and present the obtained results in Section 4. Section 5 concludes this paper by discussing these results and providing some directions for future research.

2.0 Related work

Although the notion of "term" is not yet clear, we can cite a general definition as follows: a term is "a surface representation of a specific domain concept" (Jacquemin 1997, 9). Recent research proposed to use the termhood and unithood as properties to recognize terms. According to Pazienza et al. (2005, 1), the termhood "expresses how much (the degree) a linguistic unit is related to domain-specific concepts." Mai (2008, 20) defines a domain as follows:

> An evolving and open concept that will develop as the concept is used and applied in research and practice. [T]he concept is [here] used to refer to a group of people who share common goals. A domain could, for instance, be an area of expertise, a body of literature, or a group of people working together in an organization.

According to Hannan et al. (2007), a domain is a culturally bounded segment of the social world containing producers/products, audiences, and a language that tells to whom these distinctions apply and what they mean. From these definitions, we can conclude that a domain is an area of knowledge composed of a set of related items (products). It corresponds to a common interest shared by a social community (producers and audiences having a common set of perceptions, interests, beliefs, activities, values, etc.). This community shares also a set of concepts and a terminology defined by the consensus of its members. According to Spradley (1979), a domain is defined by a

cover term (which specifies the category of the cultural knowledge), a set of included terms, semantic relationships between included terms and between the cover term and the included terms, and the means to define boundaries (criteria to decide whether an item belongs to the domain).

The unithood "expresses strength or stability of syntagmatic collocations" (Pazienza et al. 2005, 5). It concerns terms which are composed of more than one word. Multi-word expressions (MWEs) can be defined as "idiosyncratic interpretations that cross word boundaries" (Attia 2008, 71). To be considered as a MWE, a sequence of words should fulfill syntactic and semantic conditions. Attia (2008, 72) defines many properties of MWEs, such as lexogrammatical fixedness (i.e., the expression is rigid or frozen) and single-word paraphrasability (i.e., the expression can be replaced by a single word). However the main property that distinguishes these expressions is non-compositionality, which means that we cannot derive the meaning of the expression from the meanings of its components. In other words, "a multiword is a succession of words whose sense taken as a whole differs from the sum of the senses of its single words" (Martínez-Santiago et al. 2002). For example "book cover" is a compositional expression. Nevertheless, "kick the bucket" is a non-compositional expression, because its meaning (i.e., "die") is not related to any of its constituents.

Although it is difficult to decide (or to compute a binary value of) the compositionality of a given term, only non-compositional expressions are considered as eligible MWEs. However Attia (2008, 74) argues that it is possible to accept conventionalized or institutionalized expressions; these expressions "have come to such a frequent use that they block the use of other synonyms and near synonyms." We think that such expressions are useful in the context of IR tasks because they constitute good candidates for document indexing and querying. We also extract other types of expressions useful for ontology construction. Let's consider the example of the following two expressions: "اللَّبَنُ الحَار" (Al~albanu AlHaAr: the hot milk) and "المَاءُ الحَار" (AlmaA'u AlHaAr: the hot water) extracted from a corpus talking about drinks. The two heads "لَبَن" (laban: milk) and "ماء" (maA': water) represent specific domain concepts. However, the two expressions are compositional. Besides, they are neither conventionalized nor institutionalized. Nevertheless, it is useful to extract these expressions because we can infer a link between the two heads which share the same expansion ("حار": HAr, hot).

Finally, MWEs may be categorized as idioms (e.g., down the drain), phrasal verbs (e.g., rely on), verbs with particles (e.g., give up) compound nouns (e.g., book cover) and collocations (e.g., do a favor) (Attia 2008). As previously explained, our work will be limited to compound nouns. However, we do not adopt Attia's (2008, 80) definition, which considered that "a compound noun can be formed by a noun optionally followed by one or more nouns optionally followed by one or more adjectives." In fact, Arabic compound nouns are noun phrases having complex structure which should be defined more precisely according to Arabic grammar (cf. section 2.1.2).

To summarize, we extract two types of units. On the one hand, we extract simple nouns (constituted of only one word). We call "simple term" a simple noun eligible as far as termhood is concerned. On the other hand, we handle compound nouns which are noun phrases composed of more than one word and eligible in terms of unithood and termhood. This category contains non-compositional expressions and compositional ones that may be useful for indexing and querying. In the following, we call such units multiword terms (MWTs). In the remainder of this paper, simple and MWTs will be called "Domain Relevant Terms" (DRTs). The set of DRTs constitute the "Domain Terminology" (DT). Also, we extract noun phrases which head a DRT. These expressions will help infer links between DRTs.

In this context, we study the characteristics of the Arabic language which influence DRT extraction (cf. section 2.1) and existing approaches which dealt more or less with this problem. These approaches are often classified into two main categories (Pazienza et al. 2005). From one side, linguistic approaches exploit morphologic, syntactic, or semantic information implemented in language-specific rules or programs (cf. section 2.2). From the other side, statistical approaches make use of association measures exploiting frequency (cf. section 2.3). Finally, hybrid approaches try to combine linguistic and statistical techniques to recognize terms (cf. section 2.4).

2.1. Characteristics of the Arabic language

Arabic texts are ambiguous at several levels of analysis. This section focuses on problems related to terminology extraction at the morphologic and syntactic levels. Nevertheless, ambiguities in these levels influence the semantic level and consequently the whole process of ontology building.

2.1.1. The morphologic level

The Arabic language is agglutinative, derivational, and inflectional. For example, the term "وضوء" (wDw') may be analyzed as "وُضُوء" (wuDuw': ablution),"وَضُوء" (waDuw': water for ablution) or "ضَوْء" (Dw': light). In this example, the letter "و" is interpreted either as a conjunction or as the first letter of the lemma. Even in the second case, we obtain two possible lemmas diacriticised differently. In fact, the main source of ambiguity is the lack of diacritics in most existing Arabic texts. Morphological ambiguities make it difficult to extract simple terms because for each word corresponds many possible lemmas.

To reduce morphologic ambiguity, existing approaches which deal with the Arabic language are context based. Let's suppose that an entity has several possible morphologic solutions. The first step is to associate to each interpretation one or more contexts by training in a labeled corpus. In a second step, one can try to disambiguate the entities of a test collection by comparing the new contexts to those learned in the first step. This approach was implemented, for example, for POS (Part Of Speech) tagging (Diab et al. 2004) and for full morphologic analysis (Habash et al. 2009).

2.1.2. The syntactic level

There are many sources of syntactic ambiguity in the Arabic language. We can identify two types of ambiguities which influence terminology extraction. On the one hand, Arabic has a relatively free word order. For example the noun phrase "الأَكْلُ في الْبَيْتِ" (Alakolu fy Albeyti: eating in the house) may be written " في الْبَيْتِ الأَكْلُ" (fy Albeyti Alakolu: in the house, eating). On the other hand, Arabic nouns can take the role of a verb, a preposition and adverb, or an adjective. For example, the noun "البَحْث" (AlbaHth) in the sentence " أَثْمَرَ البَحْثُ عَنْ نَتَائِجَ مُثْمِرَةٍ" (Athmara AlbaHothu En nataAija muthmira: The research brought promising results) accomplishes a nominal function. However, it is considered as a verbal noun in the following sentence: "حاوَلَ البَحْثَ عَنْ حَلٍّ آخَرٍ" (HAwala AlbaHtha En Hal Akhar: He tried searching for another solution).

Syntactic ambiguities influence MWT extraction, as it is hard to identify the valid noun phrases in a sentence having many parse trees. Since MWTs have a great role in this process and, being interested in compound nouns, we start by recalling the categories of Arabic noun phrases. A noun phrase (NP) is a phrase containing a head, which is a noun or a pronoun, and, optionally, an expansion which constitutes a set of modifiers. NPs apply to syntactic rules of the language. Hence, a NP may be a unique word (a simple noun) or a composite expression. The head and the expansion are related by a syntactic relation. As detailed in Bounhas and Slimani (2009b), Arabic grammar distinguishes five types of NPs: nominal constructs (NC) (المُركّب الإضافي), adjectival phrases (AP) (المُركّب النعتي), prepositional phrases (PP) (المُركّب الحرفي), conjunctive phrases (CP) (المُركّب العطفي), and complex noun phrases (CNP) (i.e., expressions linked two or more prepositions and/or conjunctions).

2.2. Linguistic approaches

We can distinguish three main steps in a pure linguistic approach:

Parse the corpus: linguistic tools are used to tokenize the corpus. At least POS of the words are identified,

Extract candidate terms using grammar rules implemented as patterns or parsers. In this step, beginning candidate terms are mostly identified with noun phrases (Pazienza et al. 2005).

Apply filters to refine the terminology: for example by eliminating stop words, words or collocation of very common usage in language (e.g., this thing).

As example of linguistic approach applied to Arabic language, Attia (2008) presented a pure linguistic analyzer for handling MWTs. The input is a lexicon of MWTs constructed manually. Then, his system tries to identify other variations using a morphologic analyzer, a white space normalizer and a tokenizer. Precise rules take into account morphologic features such as gender and definiteness to extract MWTs. The MWTs structures are described as trees that can be parsed to identify the role of each constituent. The goal of Attia (2008) is to perform syntactic parsing and deal with linguistic ambiguities independently from the intended application or domain.

2.3. Statistical approaches

These approaches make use of statistical measures to evaluate the termhood and the unithood (cf. Pazienza et al. [2005] for description and formulae). Measures that weigh termhood are mainly based on frequency. One may assume that the more frequent a term in a document or in a corpus, the more it represents its subject. Even when combined with linguistic filters, this approach generates non-relevant candidate terms. To solve this problem, one may use TF-IDF (Salton

and McGill 1986). An example of approach employing this measure for the Arabic language is presented by Al-Qabbany et al. (2009).

MWT may be weighed in terms of termhood using the same measures. However, we need other statistical measures to evaluate the unithood. The state-of-the-art measures compute the degree of the dependency between the components of the MWT (Martínez-Santiago et al. 2002; Pazienza et al. 2005). Some of these measures were applied for the Arabic language (Boulaknadel et al. 2008; Pinto et al. 2007).

2.4. Hybrid approaches

Pure linguistic approaches are unable to give a clear definition of termhood. Statistical approaches "are unable to deal with low-frequency of MWTs" (Boulaknadel et al. 2008, 1). To avoid the weaknesses of the two approaches, a commonly recognized solution is to combine statistical calculus and linguistic knowledge. In these approaches, linguistic analysis is performed before applying statistical filters to select all linguistic admissible candidates. The accuracy of statistical measures increases because they are applied to linguistically justified candidates. Hybrid approaches may be improved by exploiting contextual information. The idea consists of using statistical measures to compute the correlation between a term and its context (Missikoff et al. 2003).

As far as Arabic language is concerned, Boulaknadel et al. (2008) presented a hybrid approach to extract MWTs from Arabic documents. They defined patterns using the POS to select candidate terms. After that, candidate terms were ranked using statistical measures. First, the approach did not include a morphologic analyzer. The integrated POS tagger (Diab et al. 2004) is unable to separate affixes, conjunctions, and some prepositions from nouns and adjectives. Second, POS tagging does not consider many features while defining MWT patterns. For example, it is not possible to impose constraints regarding the gender and/or the number of the MWT constituents. Third, this approach does not recognize the internal structure of MWTs. As previously explained, the Arabic language defines different roles of MWT constituents. Fourth, experiments were performed on only one domain, which means that the authors considered only the unithood of terms.

3.0 A hybrid approach for Arabic terminology extraction

Existing approaches on Arabic NLP and terminology extraction dealt with many steps of this process. Some researchers adopted for a purely linguistic approach for parsing and disambiguating Arabic texts (Attia 2008). Others developed statistical context-based approaches for morphologic and POS disambiguation (Diab et al. 2004; Habash et al. 2009). These works considered only the first step required for the terminology extraction process by developing NLP tools. Consequently, they are not applied to evaluate termhood or unithood. On the other side, some approaches which tried to weigh terms lack sophisticated NLP tools to extract important morphologic features and recognize the internal structure of MWTs (Boulaknadel et al. 2008). The weakness of the linguistic parsing step produces an ambiguous list of terms. For example, in Al-Qabbany et al. (2009), we find in the same cluster the words "سعودي" (a saoudian) and "السعودي" (the saoudian). Besides, there is a need to consider both termhood and unithood. These two dimensions should be taken into account early in the disambiguation step. In fact, choosing a morphologic or a syntactic solution means evaluating all the possible solutions.

Based on this discussion, we conceive a hybrid approach for Arabic terminology extraction which stands out by the following aspects. Firstly, we perform full morphosyntactic parsing of corpora. At the morphologic level, we integrate MADA, which is a linguistic tool designed to perform morphologic analysis, disambiguation and POS tagging in one fell swoop (Habash et al. 2009). At the syntactic level, we reuse a tool developed by Bounhas and Slimani (2009b). It is a shallow parser which identifies the type of each NP (i.e., adjectival, prepositional, and so on), its structure, and the roles of its constituents (e.g., "المضاف": annexed noun and "المضاف اليه": noun to which we annex).

Secondly, we use many specific-domain corpora in order to evaluate termhood besides unithood. Thirdly, we use statistical measures to weigh the two dimensions. These measures are used both for disambiguation and for DRT recognition. Consequently, we do not make a distinction between the two steps. Fourthly, the concept of relevance is not related to the distribution of terms in corpora as in TF-IDF but to complex contextual information. In our case, ambiguity resolution and domain relevance computing are seen as IR tasks where we choose the best solution (s) according to many contextual constraints (the query).

To perform this task, we have to organize knowledge present in documents by means of i) indexing models and ii) possibilistic networks which encode contextual relations (cf. section 3.1). The process of terminology extraction consists of a learning step allowing to capture initial knowledge required for relevance evaluation (cf. section 3.2) and an inference step where noun phrases are weighted (cf. section 3.3).

3.1. Knowledge modeling

Our model is inspired from possibility theory, which represents knowledge by possibilistic networks. In such networks, we define two types of edges which correspond respectively to structural contextual relations and syntactic contextual relations (cf. sections 3.1.2 and 3.1.3). The edges are weighted by the frequencies of terms in the corpus. As we explain in section 3.1.1, the frequencies may be computed according to a quantitative or a qualitative approach.

3.1.1. Quantitative versus qualitative indexing

A document analyzer (Bounhas and Slimani 2009a) is used to extract the structure of documents (i.e., the hierarchy of titles and section headings). It generates as output a list of fragments with corresponding levels in the hierarchy. If a document contains M levels, the head node(s) (e.g., the main title) is (are) assigned level M. Leaf nodes (paragraphs) are assigned level one.

Within the quantitative approach, the number of occurrence of the term t_i in the document D_j is given by:

$$Occ(t_i, d_j) = \sum_k occ(t_i, nd_k) \quad (1)$$

The value $occ(t_i,nd_k)$ is the count of the term t_i in the node nd_k.

Within the qualitative approach, the number of occurrences is computed as follows:

$$Occ(t_i, d_j) = \sum_k occ(t_i, nd_k) * level(nd_k) \quad (2)$$

Where $level(nd_k)$ is the level of nd_k in the structure of the document. With this formula, we assign greater importance to terms appearing in the head nodes than those contained in paragraphs.

In both the two cases, we compute the frequency of t_i in D_j as follows:

$$Freq_{ij} = Occ(t_i, d_j) / \sum_i Occ(t_i, d_j) \quad (3)$$

3.1.2. The structural contextual relations

The structure of a document constitutes important contextual information. We assume that the title of a composed node defines a structural context for its sub-nodes. Terms which occur in the title of a node are related to terms of its children as follows:

$$\begin{aligned} &\forall nd_i \in d, \forall nd_j \in d, path(nd_i, nd_j), level(nd_i) > level(nd_j) \\ &\forall t_i \in nd_i, \forall t_j \in nd_j, t_i \neq t_j \Rightarrow \\ &R(t_j, [Sup, t_i]) = Freq(t_j, nd_j) / (level(nd_i) - level(nd_j)) \end{aligned} \quad (4)$$

This formula considers a couple of nodes (nd_i, nd_j) which belong to a document d ($nd_i \in d$, $nd_j \in d$). The node nd_i should be one of the parents of nd_j in the structure of the document. This means that a path exists between nd_i and nd_j ($path(nd_i, nd_j)$) and that nd_i is in a higher position compared to nd_j ($level(nd_i) > level(nd_j)$). In this case, we link any two different terms t_i and t_j ($t_i \neq t_j$), which correspond respectively to the nodes nd_i and nd_j ($t_i \in nd_i$ and $t_j \in nd_j$). The edge is labeled "Sup" which stands for "Superior." This means that the term t_i is the superior of t_j or in other words, the sense of t_i generalizes the sense of t_j.

The relation has a weight equal to the frequency of the term t_j in the child node $Freq(t_j,nd_j)$, divided by the difference of level between the two nodes. This means that terms which belong to the direct children of a node will have a greater weight than terms that occur in their descendants. If we take randomly two terms, they may appear in many relative positions with different paths. In this case, we compute an average value of the "Sup" relations of all these occurrences. This kind of relation will be useful to compute the termhood of terms (cf. section 3.3.1). Indeed, we will choose the morphosyntactic solutions which are more closely correlated with their superiors.

3.1.3. The syntactic context

Given a MWT, we assume that each of its components constitutes a context for the other. Terms are linked based on the structure of MWTs. We distinguish two families of syntactic relations. On the one hand, conjunctive NPs and some NPs containing composite syntactic relations link entities in a symmetric manner. In this case, the MWT (T) is composed of two terms (t_1 and t_2) linked by a symmetric relations (sy). We compute contextual relations as follows:

$$\forall T = (t_1, t_2, sy), R(t_1, [sy, t_2]) = R(t_2, [sy, t_1]) = Freq(T) \quad (5)$$

This formula defines a contextual relation (R) which links the first term (t_1) to a context composed of the symmetric relation (sy) and the second term t_2 ([sy, t_2]). In the same manner, we link t_2 to [sy, t_1]. The weight of the two relations is equal to the frequency of the MWT in the corpus ($Freq\ (T)$).

On the other hand, non symmetric NPs are composed of a syntactic relation (ns), a head (h) and an expansion (e).

$$\forall T = (e, h, s), R(h, [ns\_expansion, e]) = R(e, [ns\_head, h]) = Freq(T) \quad (6)$$

In this case, we consider that the expansion (e) appears in a context composed of a non-symmetric relation in head (ns_head) and the head (h). In the same manner, the head (h) appears in a context composed of a non symmetric relation in expansion ($ns\_expansion$) and the expansion (e). The two relations have a weight equal to the frequency of the MWT ($Freq(T)$).

These types of relations are useful for syntactic disambiguation as we explain in section 3.3.2. Indeed, a composite NP is chosen if each of its components is correlated with the other based on frequencies we are defining in these formulae (5 and 6).

3.2. Knowledge learning

Initially, contextual relations are computed from the non ambiguous elements of all sentences in the corpus. Also, titles and subtitles of the documents are manually disambiguated. In fact, their terms represent a small percentage in terms of quantity compared to the size of the corpus, but they are the most important entities which reflect the sense of documents.

Each contextual relation is composed of a term (t_i) and a context (c_j). The latter is constituted by a relation (which can be of the form *sy*, *ns_head*, *ns_expansion* or *Sup*). The contextual relations are seen as a possibilistic network which links terms to their contexts. The graph structure encodes dependence relation sets just like Bayesian nets (Benferhat et al. 2002).

Let us take the example of the document entitled "الزواج" (AlzwAj: marriage) and already disambiguated. (cf. figure 1). Let us also consider that the whole document contains 100 terms. The node N1 entitled "لباس العرس" (lbAs AlErs: clothes of wedding) contains 20 terms. The term "لباس" (lbAs: clothes) occurs twice in N1, while the terms "الرجل" (Alrjl: the men) and "لباس الرجل" (lbAs Alrjl: clothes of the men) appear only one time in the document.

Figure 1. Example of disambiguated Arabic document and its translation.

We compute the frequencies of terms within the quantitative and qualitative approaches as in Table 1.

Frequency	Quantitative approach	Qualitative approach
Freq ("لباس", N1)	(1+1)/20 = 0.1	(2*1+1)/20 = 0.15
Freq ("عرس", N1)	(1+1)/20 = 0.1	(2*1+1)/20 = 0.15
Freq ("لباس العرس ", N1)	1/20 = 0.05	(1*2)/20 = 0.1
Freq ("رَجُل", N1)	1/20 = 0.05	1/20 = 0.05
Freq ("لباس الرَجُل ", 1)	1/20 = 0.05	1/20 = 0.15
Freq ("لباس", D)	(1+1)/100 = 0.02	(2*1+1)/100 = 0.03
Freq ("عرس", D)	1/100 = 0.01	(1*2)/100 = 0.02
Freq ("لباس العرس ", D)	1/100 = 0.01	(1*2)/100 = 0.02
Freq ("رَجُل", D)	1/100 = 0.01	1/100 = 0.01
Freq ("لباس الرَجُل ", D)	1/100 = 0.01	1/100 = 0.01

Table 1. Frequencies of terms for the document of figure 1.

We remark that the superiority relation ("Sup") between "زواج" (zwAj: marriage) and "لباس" (lbAs: clothes) occurred twice. That's why we computed the average between the weights of the two occurrences. In the last four lines of the table, "NC" stands for "nominal construct." The initial contextual relations and possibility distributions are used to treat the remaining sentences of the corpus. They are updated incrementally as far as these sentences are disambiguated.

Figures 2 and 3 represent the quantitative and qualitative networks learned from this document.

The graph represents contextual knowledge by means of weighted edges. Indeed, for each edge, the source represents a context for the destination. In these figures, the dashed lines correspond to superiority relations. The edges of this type may be seen as a tree where the most generic term is in the root (in this case it is "زواج" (zwAj: marriage)). The con-

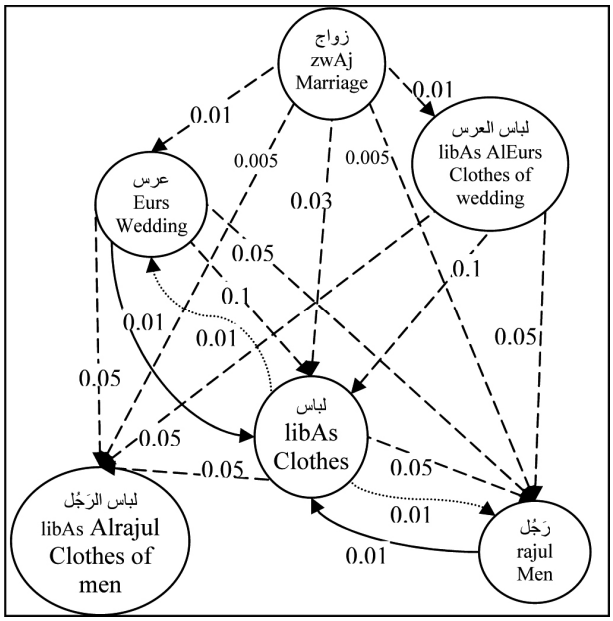


Figure 2. The qualitative network of contextual relations extracted from the document of figure 1.

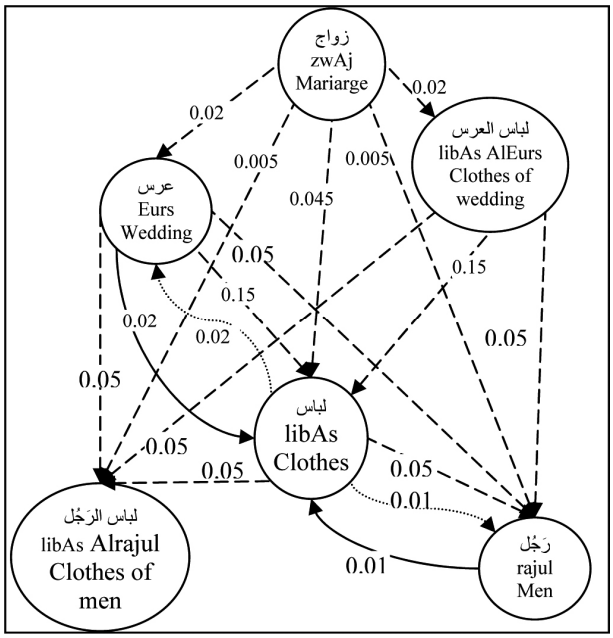


Figure 3. The qualitative network of contextual relations extracted from the document of figure 1.

tinuous and dotted lines represent “NC_head” and “NC_expansion.” The weights of the edges represent possibility distributions which are equal to the frequencies computed in table 1. For example, we note in figure 2 that π([Sup, "عرس"] |"لباس") = 0.1 which means that the term "لباس" (lbAs: clothes) appears in a context composed of the “Sup” relation and the term "عرس" (Eurs: wedding) with a weight equal to 0.1.

3.3 Knowledge inference

The contextual knowledge encoded in possibilistic networks is exploited to disambiguate the remaining nominal phrases and to evaluate their termhood and unithood in order to compute the domain relevance. Before we present our formulae illustrated with examples, we recall the matching possibilistic model used to compute the relevance of morphosyntactic solutions.

This model was initially proposed in the field of information retrieval. We suppose that there is a query Q composed of a set of items which represent constraints. We will take the general case where these items are weighted. We have:

$$Q = [(t_1, w_1)\ (t_2, w_2)\ \ldots\ (t_n, w_m)]$$

Where w_i is the weight of the term t_i.

The degree of possibilistic relevance (DPR) of a document (D_j) given the query (Q) is computed by the two measures: possibility (Π) and necessity (N).

$$DPR(D_j) = \Pi(D_j|Q) + N(D_j|Q) \quad (7)$$

According to Elayeb et al. (2009), $\Pi(D_j|Q)$ is proportional to:

$$\Pi'(D_j|Q) = Freq_{1j} * w_1 * \ldots * Freq_{mj} * w_m \quad (8)$$

The necessity of D_j for the query Q, denoted $N(D_j|Q)$, is computed as follows:

$$N(D_j|Q) = 1 - [\ (1 - \phi_{1j}/w_1) * \ldots * (1 - \phi_{mj}/\ w_m)] \quad (9)$$

Where:

$$\phi_{ij} = Log_{10}(|D|/nD_i) * (Freq_{ij}) \quad (10)$$

In this formula, $|D|$ is the number of documents. nD_i is the number of documents *containing* the term t_i (i.e., $Freq_{ij} > 0$).

In our case, each term of the query is a contextual constraint represented by a relation and a term (e.g., [Sup, "عرس"]). The documents are the morphosyntactic solutions to be weighted (e.g., "لباس" (lbAs: clothes)). The frequencies are the weights of edges linking terms in the possibilistic network.

3.3.1 Termhood evaluation

This measure weighs a candidate term according to the structural context. Given a lemma of a simple noun or a composite NP which appears in a given node (n), a query (Q) is composed of all the terms which appear in the path linking n to the root. These terms of the query are weighed according to the difference of level between the corresponding nodes (cf. section 3.3.4 for an example of query). The termhood of a term T is given by:

$$\text{Termhood ("T")} = DPR(T \mid Q) \quad (11)$$

3.3.2 Unithood evaluation

This measure is used to evaluate NPs by computing the degree of dependency between their constituents. Given a candidate NP (T) composed of two terms (t_1 and t_2) and a syntactic relation (s), we compute its unithood as follows:

$$\text{Unithood}(T) = \begin{cases} DPR(t_1 \mid [s, t_2]) * DPR(t_2 \mid [s, t_1]) \text{ if } s \text{ is symmetric} \\ DPR(t_1|[s\_expansion, t_2]) * DPR(t_2|[s\_head, t_1]) \text{ otherwise} \end{cases} \quad (12)$$

This measure considers that the two constituents are linked if each of them is relevant for the other. That's why we compute the product of the two relative DPRs.

3.3.3 The possibilistic domain relevance

The possibilistic domain relevance (PDR) of a simple noun is equal to its possibilistic termhood.

PDR(t) = termhood(t)
The PDR of a composite NP is equal to the product of the two dimensions:
PDR(t) = termhood(t) * unithood(t)
Terms which have a non null DPR are considered as DRTs.

3.3.4 Example of disambiguation

Let us consider the example of the document in figure 4. It is the document in figure 1 to which we added the word "المزخرف" (Almzxrf). We consider that the expression "لباس الرجل المزخرف" (lbAs Alrjl Almzxrf) is ambiguous. To simplify the calculus, we assume that this word has only one possible lemma (i.e., "مُزَخرَف" (muzaxraf: decorated)). In this case, we do not know if this adjective is linked to the word "الرجل" (Alrjl) or the expression "لباس الرجل" (lbAs Alrjl).

الزواج **لباس العرس** .....لباس الرجل المزخرف.....	**The marriage** **Clothes of wedding** .....decorated clothes of the men .....

Figure 4. Example of ambiguous document and its translation.

Morphological disambiguation: we disambiguate the word "الرجل" (Alrjl), which has two possible lemmas (e.g., "رَجُل" (rajul: men) and "رِجْل" (rijl: foot)). We use the structural information through the following query:

Q = ([Sup,"لباس "], 1) ([Sup,"عرس"], 1)
([Sup,"لباس العرس "], 1) ([Sup,"زواج"], 0.5) (13)

The weight of the term "زواج" (zwAj: marriage) in this query is 0.5 because the difference of level between the two nodes is 2. We compute the DPR of each solution employing the weights of the edges of the possibilistic network. By applying formula 8, we obtain:

Π("رَجُل"|Q) = π([Sup,"لباس"]| "رَجُل")* 1 * π([Sup, "عرس"]| "رَجُل")* 1 * π([Sup, "لباس العرس"]| "رَجُل")* 1 * π([Sup,"زواج"]| "رَجُل")* 0.5 = 0.05*1 * 0.05*1 *0.05*1 *0.05*0.5 = 0,175

According to (9), we have:

N("رَجُل"|Q) = 1- [(1- ϕ_{1j}/1)* (1- ϕ_{2j}/1)* (1- ϕ_{3j}/1)* (1- ϕ_{4j}/0.5)] = 1- [(1- 0.015/1)* (1- 0.015/1)* (1- 0.015/1)* (1- 0.015/0.5)] = 0.073

According to (11), we obtain:

Termhood ("رَجُل") = DPR ("رَجُل") = 0.175+0.073 = 0.248

In the same manner, we have:

Π("رجُل"|Q) = 0
N("رجُل"|Q) = 0
Termhood ("رجُل") = DPR ("رجُل") = 0

In this case, the possibilistic calculus allowed us to select the correct lemma for the word "الرجل".

Syntactic disambiguation: for the expression " لباس الرجل المزخرف" (lbAs Alrjl Almzxrf), we have to decide whether we should link the word "الرجل" (Alrajul: the men) to the word "لباس" (lbAs: clothes) (i.e., we obtain a nominal construct) or to the word "المزخرف" (Almuzaxraf: decorated) (i.e., we obtain an adjectival phrase). These two relations are non-symmetric.

As far as termhood, we obtain the same results as in morphologic disambiguation. That is:

Termhood ("لباس الرَجُل") = 0.248
Termhood ("الرَجُل المُزَخرَف") = 0

According to (12), we have:

Unithood ("لباس الرَجُل") = DPR("رَجُل" |[NC_head, لباس])*DPR (لباس|[NC_expansion, "رَجُل"])

DPR("رَجُل" |[NC_head, لباس]) = Π("رَجُل" |[NC_head, لباس])+N("رَجُل" |[NC_head, لباس]) = 0.01+0 = 0.01

DPR(لباس|[NC_expansion, "رَجُل"]) = Π(لباس | [NC_expansion, "رَجُل"])+N(لباس|[NC_expansion, "رَجُل"]) = 0.01+0 = 0.01
Unithood("لباس الرَجُل ") = 0.0001

In the same manner, we have Unithood ("الرَجُل المُزَخرَف") = 0

Finally, we have: PDR("٠.١٩٠١ = ("لباس الرَجُل and PDR("٠ = ("الرَجُل المُزَخرَف. As a result, we select the correct solution.

4.0 Experimental results

The general context of our work is a project which aims to organize documents of Arabic stories as socio-semantic maps. In this work, we are interested in the semantic axis. Our experiments in this paper constitute the first step toward the semantic representation of Arabic stories. Section 4.1 gives further information about this corpus. In section 4.2, we present our methodology of evaluation which consists of two methods of validation. We apply these methods to our corpora in section 4.3 and 4.4, respectively.

4.1. The corpus

The corpora used in the experiments are constituted from six encyclopedic books of Arabic stories grouped by theme. Story collectors grouped stories which correspond to the same domain of interest in the same chapter to facilitate their study and interpretation. Because of this structure, these books have been the subject of many works in computer and information sciences. They were studied in terms of reliability (Ghazizadeh et al. 2008; Bounhas et al. 2010). Being organized by theme, they constitute a good corpus for testing classification and clustering approaches (e.g., Al-Kabi and Al-sinjilawi 2007). They were also exploited as a corpus for testing IR systems (e.g., Harrag et al. 2009).

We can classify the knowledge organization manner in books of Arabic stories as “rationalist” since the collectors were based on a logical thematic division (Mai 2008). However, there are some differences among the classifications of the different books. Even so, we can distinguish a set of bounded domains of interest. We compile a consensual classification from the titles of chapters of the different books which constitute cover terms. Nevertheless, we preserve the internal classification of chapters of different books.

Consequently, the stories belonging to the same domain of interest may be classified into sub-domains according to many points of view corresponding to the different collectors.

The whole corpus contains more than 2.5 million words and more than 95,000 fragments (titles and paragraphs). We started by analyzing the structure of these books to extract the different themes and subthemes by using our document analyzer (Bounhas and Slimani, 2009a). This paper presents experiments on three corpora corresponding to the domains of interest "marriage" ("الزَواج": AlzawAj), "drinks" ("الأَشْرِبَة": Alachriba), and "purification" ("الطَّهَارَة": AlTahAra). Table 2 presents statistics about each domain.

The size of our corpus is comparable to other research works in the field. For example, MADA was tested on a corpus composed of approximately 51 K-words. Diab et al. (2004) tested their POS tagger with 400 sentences. Manual evaluation of the output of a morphologic analyzer or a POS tagger is hard and time-consuming. Approaches which do not perform full parsing may be evaluated in larger corpora. For example, Boulaknadel et al. (2008) evaluated their MWT extractor on a corpus containing 475,148 words. Unfortunately, there are no tokenized specialized corpora for the Arabic language. Consequently, we were obliged to build our own corpus.

4.2. The methodology of evaluation

The evaluation of knowledge extraction and IR systems is based on performance metrics. Precision, recall, and F-measure are commonly used to evaluate system performance (Rosemblat and Graham 2006). Evaluation assumes that there exists an ideal set the system is supposed to retrieve. The three metrics are defined as follows. The precision is the percentage of elements retrieved by the system, which are also in the ideal set. The recall is the percentage of elements in the ideal set that were retrieved by the system. The F-measure is given by:

$$\text{F} - \text{measure} = \frac{2 * \text{recall} * \text{precision}}{\text{recall} + \text{precision}} \quad (14)$$

Because it is hard to define the ideal set, the evaluation issue is still challenging, thus limiting the development of KOS. The evaluation of these environments is necessary to validate the theoretical assumptions and the so built resources. Unfortunately, no gold standards have been developed to assess and compare different approaches in the field. Such standards may be provided directly or through validation only by a human expert (Pazienza et al. 2005). In some cases, one can find domain knowledge organized as reference lists which may be used to evaluate system performance automatically (Martínez-Santiago et al. 2002). A reference list may also be built by a human expert who examines the corpus and extracts valid elements. When reference lists are unavailable, one can opt for the validation method where an expert validates element by element the extracted ontologies (e.g., Missikoff et al. 2003; Al-Qabbany et al. 2009). This approach is time-consuming. Also, human intervention is influenced by subjectivity and personal interpretation of terms. Finally, a terminological resource may be evaluated in the context of IR tasks. In this case, the goal is to check whether the resource will improve the performance of IR systems in terms of document retrieval.

To our knowledge, no gold standards have been developed to validate Arabic terminologies in the three considered domains. That's why we were obliged to build reference lists manually. An expert analyzes the corpora starting by titles of level 1 and 2. Because many steps in this process are manual, the quality of evaluation is influenced by subjectivity.

	Drinks	Marriage	Purification	Total
Number of titles of level 1	1	1	10	12
Number of titles of level 2	200	444	745	1389
Number of paragraphs	1897	3038	6130	11065
Number of words in level 1	1 (0.003%)	1 (0.002%)	131 (0.122%)	133 (0.069%)
Number of words in level 2	1165 (3.605%)	2669 (4.965%)	3618 (3.379%)	7452 (3.859%)
Number of words in paragraphs	31154 (96.392%)	51082 (95.033%)	103309 (96.498%)	185545 (96.073%)
Total number of words	32320	53752	107058	193130

Table 2. Statistics about fragments and terms in the three corpora.

Nevertheless, we argue that the extracted lists may be used as reference models for comparing different approaches of term extraction. Even so, we do not consider these lists as an optimal means to assess our system. To avoid this impasse and improve our assessment, we evaluate the extracted terminologies in an information retrieval system. In this step, the domain terminology is considered as a query which is supposed to retrieve the domain relevant documents. The terminologies are assessed iteratively. In each iteration, the *N* top DRTs are used to query the whole corpus. We evaluate the results in terms of precision, recall, and F-measure. Both methods of evaluation are employed to compare three approaches. In the first one, we adopt the morphologic solution chosen by MADA. Then we use TF-IDF to evaluate termhood. Finally, we employ LLR to choose the syntactic solutions and evaluate unithood. This score reached the better results in other studies (Bounhas and Slimani 2009b). The second and the third approaches use, respectively, the quantitative and qualitative possibilistic settings for morphosyntactic disambiguation, termhood, and unithood evaluation. In the following sections, we present results of evaluation within the two methods designated, respectively, "expert validation" and "system validation."

4.3. Expert validation

In this method of evaluation, we compare the list of terms returned by our system to the reference list proposed by the expert. Figures 5, 6, and 7 present

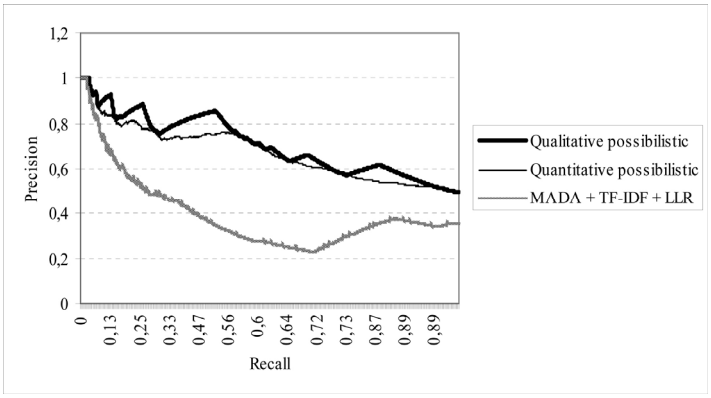


Figure 5. The curves of precision vs. recall for the domain of drinks.

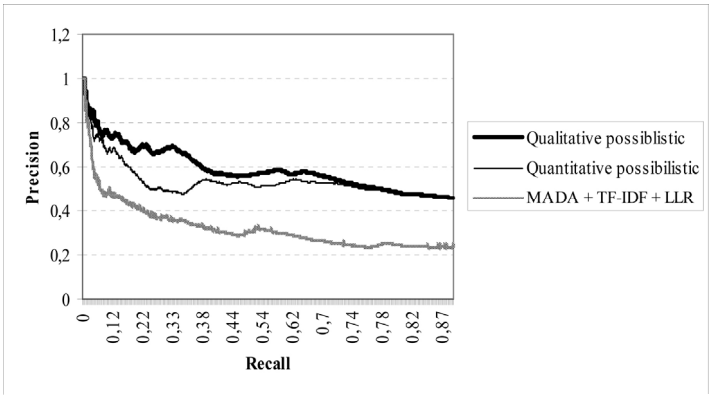


Figure 6. The curves of precision vs. recall for the domain of marriage.

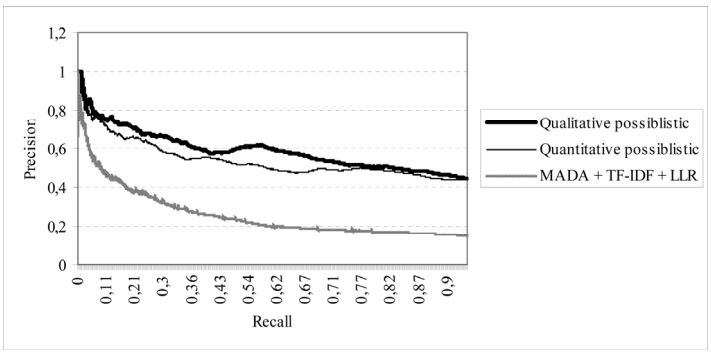


Figure 7. The curves of precision vs. recall for the domain of purification.

curves of precision versus recall for the three domains, respectively. In the three domains, the possibilistic approach improved term extraction compared to the probabilistic one (MADA + TF-IDF + LLR). This implies that domain relevance is related not only to the distribution of terms in corpora, but also to complex contextual relationships linking terms. What's more, the qualitative approach reached better results than the quantitative one. This means that terms are ranked better when their frequencies are computed according to their positions in the structure of the document.

We can study more precisely the impact of the structure by analyzing the distribution of domain relevant terms within the different levels of hierarchy. Table 3 presents the percentages of relevant terms which exist only in headings, only in paragraphs and in both for the three domains.

Domain	Only in headings	Only in paragraphs	In both
Drinks	19.83%	54.51%	25.65%
Marriage	16.13%	57.45%	26.42%
Purification	12.73%	52.08%	35.19%

Table 3. Distribution of relevant terms in the three domains.

These statistics show the importance of headings in representing the meaning of documents. Indeed, they represent only 3.927% from the number of words. However 15.52% of the relevant terms (to the three domains) exist only in these fragments. This explains the improvement realized within the qualitative approach.

We also remark that our model for organizing contextual knowledge extracts better MWTs. Indeed, structural knowledge constitutes semantic features which help in morphosyntactic disambiguation and interpretation of terms. In order to study more precisely this fact, we assessed the accuracy MWT extraction in the three domains. Our results show that using the possibilistic approach instead of MADA + TF-IDF + LLR, improves the F-measure of MWT extraction with 26.67% in average for the three domains. It reached an average value equal to 63.10%.

4.4. System validation

This method is applied twice for each domain. On the first hand, we employ all the types of terms in the queries. On the second hand, we use only MWTs. Figures 8 and 9 represent curves of F-measure versus the number of terms in the query (*N*) for the domain

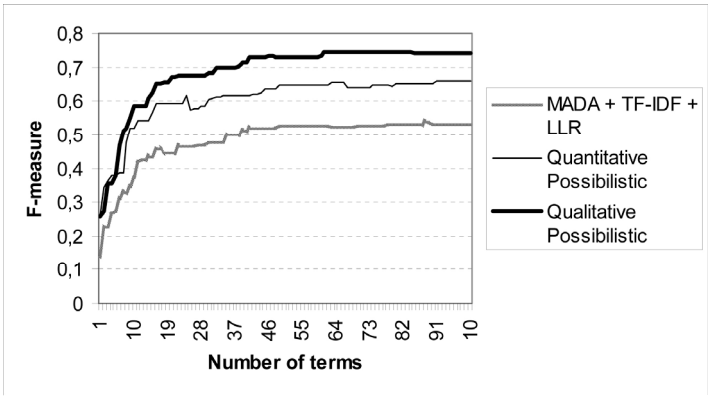


Figure 8. The curves of F-measure for the domain of purification (All terms)

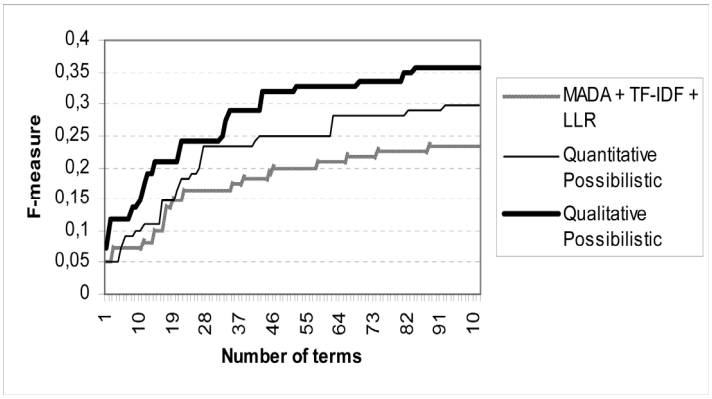


Figure 9. The curves of F-measure for the domain of purification (MWTs).

of purification within these experiments. We obtained similar curves for the two other domains. The curves show the improvement we gain by adopting the possibilistic approach compared to the one based on TF-IDF and LLR. We also see the contribution of the qualitative approach compared to the quantitative one. We compute the average of improvement of F-measure in the three domains as follows. By moving from "MADA + TF-IDF + LLR" to the quantitative possibilistic approach, we reach 8.98% and 6.87% of improvement when using all terms and MWTs, respectively. The qualitative approach performs 7.26% and 4.62% as improvement for all terms and only MWTs experimentations compared to the quantitative one. This amelioration shows, on the one hand, that our approaches extract better MWTs. On the other hand, we confirm results obtained for other languages which prove that MWTs are important entities that may be used to index and query documents (Martínez-Santiago et al. 2002).

As mean of comparison, this method of evaluation was used by Larkey et al. (2002) to assess different stemming approaches on the TREC-2001 Arabic corpus. The maximum value of F-measure of the best stemmer (light8) is about 0.43. Harrag et al. (2009), who applied their IR system in the same corpus (i.e., Arabic stories), reached an average value of F-measure equal to 0.47. In our case, F-measure reached respectively 0.88, 0.83, and 0.73 for the three domains. It is hard to compare these works because they have different goals and use different corpora and/or queries. Besides, they treated documents as a unique textual corpus while we decomposed our corpus in many specific, domain semi-structured corpora. The great improvement of the value of F-measure shown by our system is thus explained by the fact that terms which are used in the queries are already attested (according to a given measure) as DRTs.

5.0 Conclusion and future work

The experimental results show the contribution of our approaches based on complex contextual relationships compared to the state-of-the-art measures like TF-IDF and LLR used by Boulaknadel et al. (2008). This result demonstrates empirically that our model of organizing contextual knowledge based on the structure of documents has a great impact on the terminology extraction process. Consequently, the accuracy of our approach is related to the quality of the corpus. Indeed, the actual Web contains more and more semi-structured documents, while existing systems mainly focus on text collections. To generalize our results, we should apply our approach in the general context of the Web. This will allow for a better understanding of the relation between the structure and the accuracy of terminology extraction, but also to test our hypothesis in larger corpora. We should also recognize that the structure of Web documents is not necessarily hierarchical. One possible solution to be investigated is to consider types of relations other than superiority. This means that we would give a more detailed description of the structure. Weighting special parts of texts (like titles) more than other parts of text was a first approach to give them different importance. Automatic annotating techniques are useful to give more detailed structure to semi-structured documents and may be used as much by the writer or designer of a document as the reader of that document. More generally, the structure tends to highlight parts of a document. Adjoining a structure analyzer (such as the "micrological" analyzer developed by Bounhas and Slimani (2009a)) to our system should allow the recognition of the importance of particular parts of a document thanks to the interpretation of rhetoric markers as well as of spatial organizations, sizes, or styles applied on chunks of text.

Beside focusing on organizing and exploiting contextual knowledge, we were obliged to consider NLP-related tasks. The importance of NLP tools in knowledge organization tools was studied in many research works in the field (e.g., Ibekwe-Sanjuan and Sanjuan 2002; Jiang and Tan, 2010). Consequently, we investigated problems specific to the Arabic language with a view to ontology construction. It is an attempt to introduce this language into ontology engineering environments.

Finally, our tools allow us to reorganize domain knowledge in an empiricist approach (Mai 2008). The generated network encodes dependency relations between terms which may be exploited to infer semantic relations and thus build a domain ontology. In this step, distributional analysis seems to be a promising solution (Bourigault and Lame 2002; Cohen and Widdows 2009).

References

Al-kabi, Mohammed Naji and Al-sinjilawi, Saja I. 2007. A comparative study of the efficiency of different measures to classify Arabic texts. *Journal of pure & applied sciences* 4n2: 13-26.

Al-Qabbany, Abdulaziz, AbdulMalik, Al-Salman, and Abdulrahman, Almuhareb. 2009. An automatic

construction of Arabic similarity thesaurus. In Karim Bouzoubaa and Abdelfettah Hamdani ed., *Proceedings of the 3rd IEEE International Conference on Arabic Language Processing (CITALA2009)* 4-5 May 2009 Rabat, Morocco. Morocco: IEEE Morocco Section, pp. 31-36.

Attia, Mohammed. 2008. *Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation*. Ph.D. thesis, University of Manchester, Faculty of Humanities, UK.

Attia, Mohamed, Rashwan, Mohsen, Ragheb Ahmed, Al-Badrashiny, Mohamed, Al-Basoumy, Husein, and Abdou, Sherif. 2008. A compact Arabic Lexical semantics language resource based on the theory of semantic fields. In Bengt Nordström, Aarne Ranta ed., *Advances in Natural Language Processing: Proceedings of the 6th international conference on Advances in Natural Language Processing 25-27 August 2008, Gothenburg, Sweden*. Berlin, Heidelberg: Springer-Verlag, pp. 65-76.

Benferhat, Salem, Dubois, Didier, Garcia, Laurent, and Prade, Henri. 2002. On the transformation between possibilistic logic bases and possibilistic causal networks, *International journal of approximate reasoning* 29: 135-73.

Boulaknadel, Siham. 2006. Utilisation des syntagmes nominaux dans un système de recherche d'information en langue arabe. In *Proceedings of Conférence Francophone en Recherche d'Information et Applications (CORIA) 15-17 Mars 2006 Lyon, France*, pp. 341-46. Available http://bach2.imag.fr/ARIA/publisparconf.php#3

Boulaknadel, Siham, Daille, Beatrice, and Aboutajdine, Driss. 2008. A multi-word term extraction program for Arabic language. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis and Daniel Tapias ed.,. *Proceedings of the 6th international Conference on Language Resources and Evaluation (LREC) 28-30 May 2008 Marrakech, Morocco*. Paris: ELRA, pp. 1485-88.

Bounhas, Ibrahim and Slimani, Yahya. 2009a. A social approach for semi-structured document modeling and analysis. In Kecheng, Liu. ed., *Proceedings of the International Conference on Knowledge Management and Information Sharing (KMIS) 6-8 October 2009 Funchal, Madeira, Portugal*. INSTICC Press, pp. 95-102.

Bounhas, Ibrahim and Slimani, Yahya. 2009b. A hybrid approach for Arabic multi-word term extraction. In *Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE) 24-27 September 2009 Dalian, China*. Piscataway, N.J.: IEEE Computer Society, pp. 429-436. DOI: 10.1109/NLPKE.2009.5313852. Available http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5306518

Bounhas, Ibrahim, Elayeb, Bilel, Evrard, Fabrice, and Slimani, Yahya. 2010. Toward a computer study of the reliability of Arabic stories. *Journal of the American Society for Information Science and Technology* 61: 1686-1705.

Bourigault, Didier and Lame, Guiraude. 2002. Analyse distributionnelle et structuration de terminologie, Application à la construction d'une ontologie documentaire du Droit. *Traitement automatiques des langues* 43: 129-50.

Broughton,Vanda, Hansson, Joacim, Hjørland, Birger, and Lopez-Huertas, Maria J. 2005. Knowledge organization. Chapter 7 in Leif Kajberg and Leif Lorring ed., *European curriculum reflections on library and information science education.* Copenhagen: Royal School of Library and Information Science, pp. 133-148.

Cohen, Trevor and Widdows, Dominic. 2009. Empirical distributional semantics: Methods and biomedical applications Review Article. *Journal of biomedical informatics* 42: 390-405.

Diab, Mona, Kadri, Hacioglu, and Jurafsky, Daniel. 2004. Automatic tagging of Arabic text: From raw text to base phrase chunks. In Julia Hirschberg ed.,. *Proceedings of The 5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04*) 2-7 May 2004 Boston, Massachusetts, USA. East Stroudsburg, PA: Assoc. for Computational Linguistics, pp. 149-52.

Dunning, Ted. 1994. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics* 19: 61-74.

Elayeb, Bilel, Evrard, Fabrice, Zaghdoud, Montaceur, and Ben Ahmed, Mohamed. 2009. Towards an intelligent possibilistic web information retrieval using multiagent system. *The international journal of interactive technology and smart education (ITSE), Special issue: New learning support systems* 6: 40-59.

Elkateb, Sabri, Black, William J., Vossen, Piek, Rodríguez, Horacio, Pease, Adam, Alkhalifa, Musa, and Christiane, Fellbaum. 2006. Building a WordNet for Arabic. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk and Daniel Tapias, ed., *Proceed-*

ings of the 5th Conference on Language Resources and Evaluation (LREC2006) 24-26 May 2006 Genoa Italy. Paris: ELRA, pp. 29-34.

Ferret, Olivier, Grau, Brigitte, Hurault-Plantet, Martine, Illouz, Gabriel, Jacquemin, Christian, Monceaux, Laura, Robba, Isabelle, and Vilnat, Anne. 2002. How NLP can improve question answering. *Knowledge organization* 29: 135-55.

Ghazizadeh, Mehdi, Zahedi, M. Hadi, Kahani, Mohsen, and Bidgoli Minaei B. 2008. Fuzzy expert system in determining Hadith validity. *Advances in computer and information sciences and engineering* 354-59.

Habash, Nizar, Rambow, Owen, and Roth, Ryan. 2009. MADA + TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In Khalid Choukri and Bente Maegaard ed.,. *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)* 22-23 April 2009 Cairo, Egypt. Cairo: MEDAR Consortium, pp. 102-9.

Hannan , Michael T., Pólos, László, and Carroll, Glenn R. 2007. *Logics of organization theory: audiences, codes, and ecologies.* Princeton: Princeton University Press.

Harrag, Fouzi, Hamdi-Cherif, Aboubekeu, Al-Salman, Abdul Malik S., and El-Qawasmeh, Eyas. 2009. Experiments in improvement of Arabic information retrieval. In Karim Bouzoubaa and Abdelfettah Hamdani ed., *Proceedings of the 3rd IEEE International Conference on Arabic Language Processing (CITALA2009) 4-5 May 2009 Rabat, Morocc.* Rabat: Mohammadia School of Engineers, pp. 71-81.

Ibekwe-Sanjuan, Fidelia, Sanjuan, Eric. 2002. From term variants to research topics, *Knowledge organization* 29: 181-97.

Jacquemin, Christian. 1997. Variation terminologique: Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. *H.Dr. Thesis in fundamental computer science*. University of Nantes, France.

Jiang, Xing and Tan, Ah-Hwee. 2010. CRCTOL: A Semantic-based domain ontology learning system. *Journal of the American Society for Information Science and Technology* 61: 150-68.

Larkey, Leah S., Ballesteros, Lisa, and Connell, Margaret E. 2002. Improving stemming for Arabic information retrieval: Light stemming and cooccurrence analysis. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval 11-15 August 2002, Tampere, Finlande*, New York, NY, USA: ACM, pp. 275-82.

Mai, Jens-Erik. 2008. Design and construction of controlled vocabularies: Analysis of actors, domain, and constraints. *Knowledge organization* 35: 16-29.

Malaisé, Véronique, Zweigenbaum, Pierre, and Bachimont, Bruno. 2003. Vers une combinaison de méthodologies pour la structuration de termes en corpus : Premier pas vers des ontologies dédiées à l'indexation de documents audiovisuels. In Widad Mustafa El Hadi ed.,. *Actes du 4e Congrès ISKO France 3-4 july 2003 Grenoble France*. Paris: L'Harmattan, pp. 179-89.

Martínez-Santiago, Fernando., Díaz-Galiano, Manuel Carlos, Martín-Valdivia, Maite Teresa, Rivas-Santos, Víctor Manuel, and Ureña-López, Luis Alfonso. 2002. Using neural networks for multiword recognition in IR. In López-Huertas, M. J. ed.,: *Challenges in knowledge representation and organization for the 21st century: Integration of knowledge across boundaries: Proceedings of the Seventh International ISKO Conference 10-13 July 2002 Granada, España*. Advances in knowledge organization 8. Würzburg: Ergon, pp. 559-64.

Missikoff, Michele, Velardi, Paolo, and Fabriani, Paolo. 2003. Text mining techniques to automatically enrich a domain ontology. *Applied intelligence* 18: 323–40.

Pazienza, Maria Teresa, Pennacchiotti, Marco, and Zanzotto, Fabio Massimo. 2005. Terminology extraction: An analysis of linguistic and statistical approaches. In Spiros Sirmakessis, ed., *Knowledge mining series: Studies in fuzziness and soft computing*. Berlin, Heidelberg: Springer, pp. 255–79.

Pinto, David, Rosso, Paolo, Benajiba, Yassine, Ahachad, Anas, and Jiménez-salazar, Héctor. 2007. Word sense induction in the Arabic Language: A self-term expansion based approach. In Adeeb Riad Ghonaimy, ed., *Proceedings of the 7th Conference on Language Engineering, The Egyptian Society of Language Engineering 5-6 December 2007 Cairo, Egypt*. Cairo, Egyptian Society of Language Engineering, pp. 235-45.

Rodríguez, Horacio, Farwell, David, Farreres, Javi, Bertran, Manuel, Alkhalifa, Musa, and Martí, M. Antonia. 2008. Arabic WordNet: Semi-automatic extensions using Bayesian Inference. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis and Daniel Tapias ed., *Proceedings of the 6th interna-*

tional Conference on Language Resources and Evaluation (LREC) 28-30 May 2008 Marrakech, Morocco. Paris: ELRA, pp. 1702-06.

Rosemblat, Graciela and Graham, Laurel. 2006. Cross-Language search in a monolingual health information system: Flexible designs and lexical processes. In Budin, Gerhard, Swertz, Christian, and Mitgutsch, Konstantin, ed., *Knowledge organization for a global learning society: Proceedings of the Ninth International ISKO Conference 4-7 July 2006 Vienna, Austria*. Advances in knowledge organization 10. Würzburg: Ergon-Verlag, pp. 173-82.

Salton, Gerard and McGill, Michael J. 1986. *Introduction to modern information retrieval*. New York, NY, USA.: McGraw-Hill, Inc.

Souza, Renato Rocha and Raghavan, K.S. 2006. A methodology for noun phrase-based automatic indexing. *Knowledge organization* 33: 45-56.

Spradley, James P. 1979. *The ethnographic interview*, New York: Holt, Rinehart and Winston.

Zaidi, Soraya and Laskri, Mohamed Tayeb. 2005. A cross-language information retrieval based on an Arabic ontology in the legal domain. In Richard Chbeir, Albert Dipanda and Kokou Yétongnon *eds., Proceedings of the 1st International Conference on Signal-Image Technology and Internet-Based System (SITIS) November 27 - December 1 2005 Yaounde, Cameroon*. Dicolor Press, pp. 86-91.