A morpho-semantic preparation of the Arabic queries to improve the calculation of relevance in the IRS

Adil ENAANAI[†] and Aziz SDIGUI DOUKKALI[††] and El habib BELAHMER[†††]

Centre d'Etudes doctorales ST2I, Ecole nationale d'informatique est d'analyse des systèmes, Rabat, Morocco

Summary
In the information systems, the query's expansion brings more benefices in the relevant documents extraction. However, the current expansion types are focused on the retrieve of the maximum of documents (reduce the silence). In Arabic, the queries are derived in many morphosemantical variants. Hence the diversity of the semantic interpretations that often creates a problem of ambiguity. Our objective is to prepare the Arabic request before its introduction to the document retrieval system. This type of preparation is based on pretreatment which makes morphological changes to the query by separating affixes of the words. Then, present all of morphosemantical derivatives as a first step to the lexical audit agent, and check the consistency between the words by the context parser. Finally we present a new method of semantic similarity based on the equivalence probability calculation between two words.

Key words:
Relevance; Information research; Similarity; Semantic gene; Arab treatment.

1. Introduction

In the information research systems, relevance is function of the similarity degree between the query and the document. However, many functions of similarity are recently proposed. Most of these functions are based on the principle of vector distance where the meaning of words is not supported. However two words whose distance is zero, meaning they are similar. Unlike, there are words of high distance that mark the same meaning (synonyms), or words of distance equal to zero which mean several things.

Functions based on the distance vector are unable to provide the exact value of semantic similarity of terms. Withal, there are many stemming algorithms which contributing to the calculation of relevance by comparing the roots of words. That has yielded good results. But still insufficient, because there are some Arabic words whose roots are written in the same way while their meaning is different.

The difficulty of having a function of semantic similarity lies in the fact that the comparison of meaning between two words is possible only after an inclusion of a valid morphosemantical analysis. Hence, the need to properly prepare the query before being introduced to the data retrieval system.

Manuscript received April 5, 2012
Manuscript revised April 20, 2012

2. State of the Art

Most of the works on the Arabic syntactic analysis have led to the achievement of some laboratory prototypes. Indeed, to our knowledge to date, there is no parser marketed or distributed for scientific research. In the remainder of the State of the art, we present a few of the Arabic language analysis systems.

2.1 The AraParse system for syntactic analysis of the Arabic unvowled

AraParse is a system to analyze Arab texts in their vowelized, unvowelized or partially vowelized form [1]. The objective is to achieve a core of morpho-syntactic analysis system that can be reused in other applications such as information research and automatic translation.

AraParse is based on linguistic resources in wide coverage. It uses a lexicon of lemmas generated from the DIINAR.1 dictionary [3]. This glossary contains 19 6818 unvowelized bases distributed to 39 000 nominal bases, 79 818 verbal bases and 78 000 derivable bases from 20 000 verbs of DIINAR.1([2], [1]). To recognize unknown sequences or unknown words, this system uses an approximate matching technique implemented with the AGFL formalism and using the priority operator between the alternatives of a rule and regular expressions [1].

Ouersighni [1] proposed the use of AraParse detect and diagnose the faults of accord. He used the accord rules proposed by Belguith [4] in the DECORA system.

2.2 The system DECORA for detection and correction of the Arabic accord errors

In scientific research focused on the analysis of Arabic language, Belguith [4] proposed a method for detecting and correcting errors in accord. This method has been implemented in the system DECORA. It is based on syntagmatic analysis for the error detection and correction multi-criteria analysis. Extended sentence is defined as a group of one or more original sentences linked to accord between them.

Extended syntagmatic analysis operates in two stages ([5], [6]). The first step is to cut the initial phrase in initial phrases by locating the boundaries between them. This Division is guided by a set of rules using the syntagmatic borders as a means of identification of the original phrases. The second stage, allows to build the extended phrases. The constitution of these phrases is guided by indicators of surface and is based on a set of rules to locate the accord links between the original phrases. These rules allow for example to relate the possessive pronouns in the phrases to which they relate, to integrate the verbs in the phrase that contains the subject considered to relate the original phrases that represent anaphoric proposals to the phrase containing the syntactic unit to which it relates.

2.3 spoken Arabic Levantin Analyzer

Chiang [7] is interested by the analysis of the Arabic Levantin (AL) (a group of Arab dialects spoken in Syria, Palestine, Western Jordan and the Lebanon). He proposed an approach to translate the AL in Standard modern Arabic (SMA). Then link the sentence in AL to the corresponding analysis in SMA.
Note that the automatic translation is particularly difficult when there is no resource available as the parallel texts or the transfer lexicons. Thus, Chiang is primarily based on a corpus annotated from modern standard Arabic (MSA Treebank) [8] as well as a corpus annotated Arabic Levantine and more specifically that of the Jordanian dialect (i.e., TBPC Treebank [10]).
He built a lexicon contains the AL/ASM pairs of the forms of words. Also he built a synchronous grammar ASM-dialect. He assumes that each tree in the grammar of modern standard Arabic extracted from the MSA Treebank is also a tree of Levantine Arabic given syntactic similarity between the DSO and the AL.

2.4 A morphological and syntactic Arab text Analyzer

Debili Zouari [9] proposed the automatic construction of a dictionary containing all the inflected forms. This construction is made by a conjugator and a derivator.
The principle of morphological analysis is to make:

- The division of the text to graphical words.
- Research of enclitic and proclitic of the word.
- Verification (for each possible division) of the compatibility (proclitic / enclitic; enclitic / root; root / proclitic).

When consulting the dictionary, Zouari and Debili use the rewrite rules to find the "normal" form of the word.
The parsing process follows the phase of morphological analysis and related on the construction of the frequency binary and ternary matrices of precedence. These matrices are constructed from the annotated start texts "by hand" (this is the learning phase). They are then used to analyze new texts.

2.5 IRLA analyzer

The IRLA Analyzer is a queries interrogation system in the Arabic natural language [11]. It takes in input an Arabic sentence and translates it as query to run by an operating system. This parser allows to treat a subset of natural language (i.e., essentially imperative sentences), it produces a parenthesed form expressing the semantic of the query [12].
The parser can treat some simple linguistic problems (synonymy, negation, coordination). It is based on the detection of conceptual and linguistic surface indicators at the analysis.

2.6 Elliptical sentences Analyzer

In its research work on the Arabic analysis, Haddar [13] conducted a parser for the detection and resolution of the elliptical sentences in the Arabic texts. This parser is based on a method of syntactic analysis for verification of the syntactic structures of the proposals. This method uses a formal grammar rules generating verbal proposals written in Arabic. Access to these rules is coordinated with increased transitions (ATN) networks. The parser is coupled with another parser treating with semantic ellipses.

3. Contribution

The calculation of relevance in our approach is focused on semantic similarity function which gives a result as a percentage of equivalence between two Arabic words. Knowing that they are written in various derived forms, it had to begin by morphological analysis which returns the origin of the derivative in question. Therefore, the possibility of separate affixes of the word is subsequently obtained by the original non-vocalized of the word which may refer to several meanings. The probable meaning to be just, is that which is on conflict with the user profile. To filter the true meaning, we have developed an automatic profiling system that brings together user queries and implements format indexed in a database. Our approach has given a good result on the morphosemantical ambiguity. In the remainder of this article we will present the various stages of analysis that we introduced in the relevance calculation [17] [18].

3.1 The morphological analysis

After a sending of query through the meta-search engine, we get a list of results. This list is sorted according to the

relevance algorithm used in each data source (search engine). We begin firstly by reindexing of the documents founded by a semantic analysis module. Indeed, this module receives three parameters: the document, the query and the user profile. The document and the query are an affected by an in-page modification, which dissects the words to remove affixes [13]. If there are several cases (ex: بطريق = > بطريق = 'Penguin' or ب طريق = 'by road'), we tests the consistency of each derived, relatively with other words in the query, document and profile. Therefore, we accept just the possibility which is in the current context [5]. The separation of the terms of the request and the document offers more precision to the similarity of the triplet: document, request and profile, Hence, the need to have a flexible, fast and easy method [17].

Stemming used methods are unable to resolve the problem of semantics, because, to return the root of a word means that we can derive several forms to build a set of words that are not necessarily similar on the semantic meaning [6]. For example, the root of the word 'طريق' is 'طرق'. This root may take several derived forms, same as 'مطرقة', 'طريقة' which do not mean the same thing. Hence, we have to found the word origin by keeping the semantic aspect of all. The method that we have introduced is to dissect the word to draw the origin after applying a light stemming. The origin is later transformed into a singular to test its existence in the dictionary of the Arabic words (ARRAMOOZ ALWASEET dictionary). If the word exists, then we retrieve its definition and we type to construct a semantic entity (SE). These SE are used to test the consistency of the word in a text. For this, presents the following heuristic algorithm:

```
M=Set of arabic words
D=Set of the derived forms
P=Set of the prefixs
S=Set of the suffixs
Mot=The word to derive
Prefix ϵ P
Suffix ϵ S
Fonction derive(Mot, pref, suff, def)
Mot ←Singular(Mot)
IF (Mot in M) AND NOT (Mot in D)
 Add Mot to D
End If
If (Mot is begined by 'ال' ) AND (def=null) AND NOT ((Mot is ended by Suffix) OR (suff<>null))
m←Mot – 'ال'
def←'ال'
Derive(m,pref,suff,def)
End If
If (Mot is begined by Prefix) AND (pref=null) AND (def=null)
 M←Mot-Préfix
 Pref←Préfixe
 Derive(m,pref,suff,def)
End If
If (Mot is ended by Suffix) ET (suff=null) ET (def=null)
m←Mot – Suffix
suff←Suffix
Derive (m,pref,suff, def)
End If
End Function
```

Example: Word = 'بطريقنا' (En: our path)

At the first entry in the *derive* function, the word 'بطريقنا' is not found in in the Set of words M. we test if the Word 'Mot' is begun by a prefix. If yes, we remove the prefix to have the new word 'طريقنا' (En: our path) which will be introduced as a parameter to the recursive function 'derive'. At its entry, the word 'طريقنا' is undefined (نكرة) and is not in the set of words, and not starting with a prefix. Therefore, we pass to the second test on the suffix. The word ends with the suffix 'نا' (En: our), it also removed to have the newest derived word 'طريق' (En: path) which will also be introduced in the third hierarchical level of the recursive function. At its entry we test again if the word 'طريق' exists in the set of the Arabic words. Now, the word is founded and added in the set D of the derived words.

On returning to the first hierarchical level of recursion, the word 'بطريقنا' must be passed to another test of suffix. Then, we remove the suffix 'نا' to have the word: 'بطريق' (En: Penguin). This last word passes as a parameter to the function 'derive' which tests its existence in the set 'M'. We find also the word 'بطريق' in 'M' and add to the set of derived words 'D'. The following figure illustrates the changes [14].

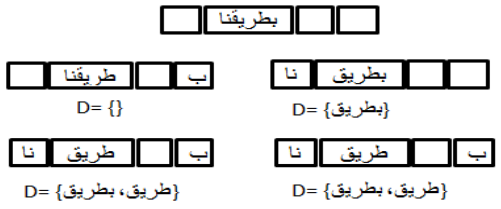

Figure 1 : The morphological analysis

In this way, the statistical parser and the profiling algorithm receive well presented data. Then, we will see how to use the morphological analysis to give valid semantic presentations [11].

3.2 Semantic gene building

The semantic gene is an object containing the information needed (from a database) to the disambiguation of the Arabic words. The construction of gene starts at the level of the morphological analysis in the determination of the origin of the word. The following diagram illustrates the format of the semantic gene [17].

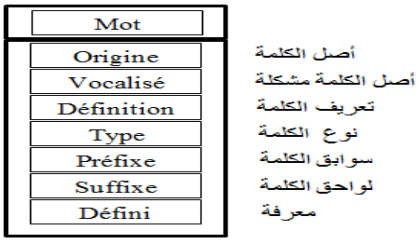

Figure 2: Structure of the semantic gene

3.3 two words queries Analysis

The expression of need is made by a multi-word query. The information system research returns a set of documents that contain all of the semantically valid sentences. These sentences include those that contain the desired word. This Word can mean several things, Hence the problem of semantic ambiguity. To reduce the effect of this type of ambiguity, we designed a semantic filtering system that recognizes the type of the word based on the rules of constitution of the Arabic sentences. Given the difficulty of semantic analysis of the Arabic sentences, we consider the case of significant sentences of two lemmas. The following table shows the different cases of a sentence of two words semantically consistent [7].

Table 2: Types of words

Object		Property		Fact	
Type	Acronyms	*Type*	Acronyms	*Type*	Acronyms
صيغة مبالغة	O_{Mob}	اسم تفضيل	P_{Taf}	مصدر	F
منسوب	O_{Man}	اسم مفعول	P_{Maf}		
جامد	O_{Jam}	اسم فاعل	P_{Fa}		
اسم مفعول	O_{Maf}	صفة مشبهة	P_{SM}		
اسم فاعل	O_{Fa}	صفة	P_S		
		منسوب	P_M		

Note: Just the types listed in the table above are considered.

Table 3: Table abstract of the synthetic mussels

Prefix of the word1	Word1	R	Word2	Example	Pattern Mot1 R Mot2	Relation between Word1 and word2
(ك، ل، و، ب)	D	∅	D	المدينة القديمة	O P	M2 is P for M1
				الإعادة البطيئة	F P	M2 is P for M1
(ك، ل، و، ب)	I	∅	D	مدينة العرفان	O O	M2 is a specification for M1
				طرح التساؤلات	F F	M1 is an action Applied on M2
				تسيير الشركة	F O	M1 is an action Applied on M2
				أرقى العائلات	P O	M1 is P for M2
				أضعف الاحتمالات	P F	M1 is P for M2
(و)	D	∅	I	المسألة صعبة	O P	M2 is P for M1
				التسارع بطيء	F P	M2 is P for M1
(ك، ل، و، ب)	I	∅	I	سنة سعيدة	O P	M2 is P for M1
				تفكير سليم	F P	M2 is P for M1
				تزويج قاصر	F O	M1 is an action Applied on M2
				أقوى رجل	P O	M1 is P for M2
				أصعب اختبار	P F	M1 is P for M2
(ك، و، ب)	D	ل	D	التعليم ل الشعب	F ل O	M1 is a F applied for M2
				الماء ل القرى	O ل O	M2 is a specification for M1
(ك، ل، و، ب)	D	ب	D	الأمر ب المعروف	F ب O	M2 is a specification for M1
…	…	…	…	…	…	…

We note that there are prohibited cases (as: " O و P"). Therefore, we have designed a set of mussels forming all possible cases of the sentences of two lemmas. This set of mussels is an array of objects where each element describes a phrase (pattern) model. The process of

correction is applied firstly on the list of the semantic entities (alimented query) of the user to remove the inconsistent morphosemantical variants [12]. Then, we send the remaining lists for contextual correction system. The latter uses the contextual corpus to refilter the list. The result is one or more lists of consistent semantic entities at the contextual level as at the semantic level. Finally, the research system is receives a suite of semantic genes containing all information that can help the extraction, selection and filtering of relevant documents [8].

There are words that can be objects or properties. Our approach supports the gene by assigning the type of the word [17] [18].

Example:

Let's say that we have the sentence ph = "الحاكم العادل" (En: Just governor).

Word1= « الحاكم » ; Word2= « العادل » ; R= « ∅ »

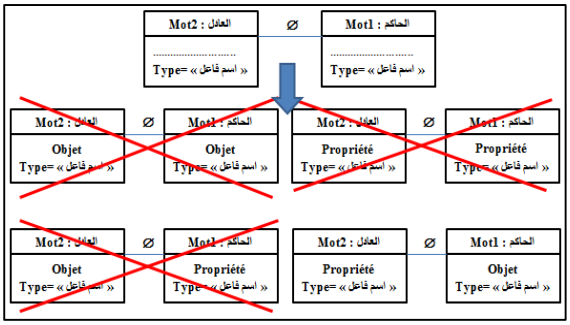

Figure 3: Two words query analysis

The table of mussels shows that the only case which exists in is: Word1 = "object"; Word2 = "property". Therefore, the sentence "العادل الحاكم" is semantically different to "الحاكم العادل", because its components are not similar. In this way, our system will be able to considerate polysemy [9].

This work is an aspect that has largely been addressed to the Latin language (English, French, ...) and even in some work for the Arabic language. Indeed, research based on the user profile to reduce noise and silence in the information research has yielded satisfactory results especially with the modeling of the user profile and the research domain with the notion of ontology. However, the ambiguity in the terms of query cannot guess the domain to choose from. Hence, we must prepare the query to reduce morphosemantical ambiguity, then guess the context from the user profile and create genes to clarify the semantic field and the context intended by the user. The following diagram illustrates the various steps of our approach [16] [17].

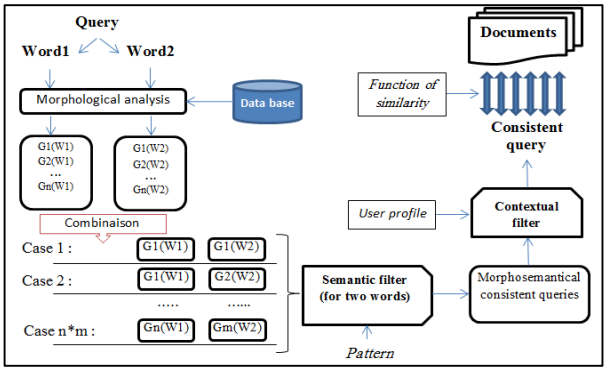

Figure 4: General diagram of approach steps

4. Evaluation

To test the effectiveness of our method, we have developed a test meta-search engine. The latter uses data sources "Bing", "Yahoo", "Yandex". Then we compared our results with Google results. We have obtained the following table after throwing 100 queries [17].

Table 4: Evaluation of the approach

	Google	Our system
Average of number of relevent liks sorted in the midst of the ten first positions	**6.47**	**8.14**

5. Conclusion et perspectives

In this article, we introduced the concept of the semantic gene that contributes to the Elimination of ambiguity in the information research systems. We also explained how to create the semantic genes from the morphological, contextual and semantic analysis and how to differentiate between homonyms. The automatic profiling is also an interesting factor to approach to the needs of users.
Our target is to automatically create semantic graphs whose semantic genes nodes are very rich in side informational data. Where each node has a context, a definition, a type of Word, a morphological form, a list of successors and a list of predecessors. Finally we wish to develop a meta-research engine which can return optimal results.

References

[1] R. Ouersighni, La conception et la réalisation d'un système d'analyse morpho-syntaxique robuste pour l'arabe : utilisation pour la détection et le diagnostic des fautes d'accord, Thèse de doctorat, université lumière–Lyon2, 2002.

[2] J. Dichy, Morphosyntactic Specifiers to be associated to arabic lexical entries-Methodological and theorical aspects, actes de la conférence ACIDCA'2000, Volume Corpora and Natural Language Processing, Monastir-Tunisie, pp. 55–60, 22–24 Mars 2000.

[3] J. Dichy, On lemmatization in Arabic : A formal definition of the Arabic entries of multilingual lexical databases, In proceedings of the Arabic Language Processing workshop, Association for computational linguistics (ACL) 39th annual meeting and 10th conference of the european Chapter, Toulouse, Juillet 2001.

[4] L. Belguith Hadrich, Traitement des erreurs d'accord de l'arabe basé sur une analyse syntagmatique étendue pour la vérification et une analyse multicritère pour la correction, Thèse de doctorat en Informatique, Faculté des Sciences de Tunis, Février 1999.

[5] L. Belguith Hadrich, A. Ben Hamadou et C. Aloulou, Using the TOPSIS multicriteria method to direct an agreement error correction process : An application to Arabic, Recent Advanced Natural Language Processing II, vol.189, pp. 105–114, John Benjamins, Amsterdam/ Philadelphia, 2002.

[6] L. Belguith Hadrich, A. Ben Hamadou, Traitement des erreurs d'accord : une analyse syntagmatique pour la vérification et une analyse multicritère pour la correction, Revue d'Intelligence Artificielle (RSTI – RIA), Editions Hermès-Lavoisier, vol. 18, N5 et 6, pp. 679–707, Décembre 2004.

[7] Chiang, , M.Diab, N. Habash, O. Rambow, S. Shareef, Arabic Dialect Parsing, In Proceedings of the European chapter of the Association of Computational Linguistics EACL, , pp. 369–376, Trento, 2006.

[8] M. Maamouri, A. Bies, T. Buckwalter, The Penn Arabic Treebank : Building a large-scale annotated Arabic corpus, In NEMLAR Conference on Arabic Language Resources and Tools, pp. 102–109, Cairo, Egypt, 2004.

[9] Debili et L. Zouari, Analyse morphologique de l'arabe écrit voyellé ou non fondée sur la construction automatique d'un dictionnaire arabe, Cognitiva, Paris, 1985.

[10] M. Maamouri, A. Bies, T. Buckwalter, M. Diab, N. Habash, O. Rambow, D.Tabessi, Developing and using a pilot dialectal Arabic Treebank, In Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06, Genoa, Italy, 2006.

[11] N. Fourati et A. Ben Hamadou, The Linguistic Portability of the Natural language Interfaces, 9th CLIN (Computational Linguistics in the Netherlands) Meeting, University of Leuven, 11 December, 1998.

[12] N. Fourati, A. Ben Hamadou et F. Gargouri, An object-orientated pivot model for the design of human-machine interfaces in natural language, 6th International Conference and Exhibition on Multi-lingual Computing Cambridge, 17–18 April 1998.

[13] K. Haddar, Caractérisation formelle de l'ellipse et processus de recouvrement de la langue arabe, Thèse de doctorat, Université de Tunis II – Faculté des Sciences de Tunis, juillet 2000.

[14] El YOUNOUSSI Y (2011), La racinisation de la langue arabe par lesautomates à états finis (AEF), 4th International Conference on Arabic Language Processing.

[15] Alrahabi M. (2004). Filtrage sémantique de textes en arabe en vue d'un prototype de résumé automatique, JEP-TALN 2004, Traitement Automatique de l'Arabe Fes

[16] Collet K (2003) Méthode du TALN, traitement automatise du langage naturel, notion de l'indexation automatique, Cours, URFIST Bretagne Loire-Atlantique.

[17] A. ENAANAI and A.DOUKKALI (2012), An hybrid approach to calculate relevance in the meta-search engines, IJSAT, volume 2, N°3. March 2012.

[18] Benlahmar H. (2006). A New Solution for Data Extraction: GENE/LONE Method, IJCSNS International Journal of Computer Science and Network Security, Vol 6, N° 7.

Adil ENAANAI was born in 1981 in Fes in Morocco took his Bachelor of mathematical and experimental sciences in 2001, he has continued his university studies at the Faculty of science ben m'sik for having an applied license in computer science. Then he went to the capital to prepare a master of computer engineering research degree. After that, he began his doctoral studies at the CEDoc ENSIAS where the subject is "the meta search in Arabic language". Also, he obtained the diploma of specialized technician in industrial automatism, and the diploma of the normal superior school in teaching pedagogy. He worked as Professor of computer science for six years which the specialty was the web development and research information on the internet.