Information Retrieval in Arabic Language

Malek Boualem (1) Ramzi Abbes (2)

(1) France Telecom Orange Labs, France
Email : malek.boualem@orange-ftgroup.com

(2) Lyon 2 University / ICAR-CNRS, France
Email : ramzi.abbes@univ-lyon2.fr

Abstract

Web search engines provide quite good results for Latin characters-based languages. However, they still show many weaknesses when searching in other languages such as Arabic. This paper discusses a qualitative analysis of information retrieval in Arabic, highlighting some of the numerous limitations of available search engines, mainly when they are not properly adapted to the Arabic language features. To support our analysis we present some results based on thorough observations about various Arabic linguistic phenomena. To validate these observations, we mainly have tested the Google search engine. Arabic information retrieval still faces many difficulties due to the Arabic linguistic features, especially its complex morphology and the absence of vowels in available documents and texts. These specificities often cause significant dissymmetry between the indexation process and the query analysis. We present in this paper some of the morphological constraints of Arabic language and we show through experimental tests how search engines deal with them. Finally this paper clearly states that information retrieval in Arabic language will never succeed without including language processing tools at all the linguistic levels (lexical, syntactic and semantic).

Keywords: Information retrieval, Arabic language, Google

1 Introduction

With 90,83% of the internet users (see Figure 1, December 2007[1]), Google is probably the most powerful search engine on the market, or more precisely the most used one, because there is correlation between these two aspects. Indeed the Google PageRank algorithms are very sensitive to the user's behaviour (Brin; Page, 1988). These algorithms balance positively or negatively web pages according to the click numbers on the corresponding links and the PageRank scores web pages according to the number of hypertext links they contain (Chen & al., 2007). This observation is also very accurate when using Google to search in Arabic language. For example, we have noticed that most of the top list answers correspond to various forums منتديات or to some other specific information sources. The Google PageRank attributes higher scores to these information sources as they contain a high number of hypertext links and also because they are more commonly used in the Arabic world instead of other information sources such as scientific papers or publications.

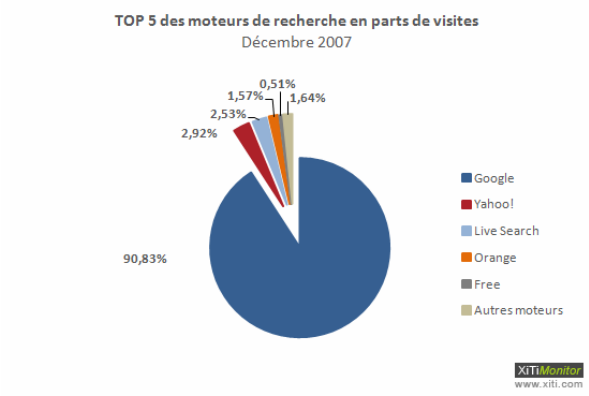

Figure 1. Top 5 of search engines based on the user's number

Most of the Arabic internet users master a second language, mainly English or French. As the information on the Web is widely available in these languages, the Arabic internet users often prefer searching in these languages rather than in Arabic. Of course consequently this situation does not help the development of Arabic information resources on the Web.

Moreover to avoid editing problems of Arabic texts on various screens and operating systems, a lot of publishers (e.g. newspapers) provide the Web with PDF documents. This situation also does not help searching information in Arabic language.

[1] http://barometre.secrets2moteurs.com/index.php/Barometre-1ere-position-xiti

2 Arabic language on the Web

2.1 Dissymmetry between indexation and query analysis processes

Searching in Arabic language meets a fundamental problem related to a certain dissymmetry between the indexation side and the query analysis side. One of the reasons is related to the Arabic vowels. For example, in the indexation process, the verb "to write" (كَتَب) together with the noun "books" (كُتُب) are indexed under the same entry كتب, since they probably do not have vowels in the Arabic original text. This problem concerns most of the Arabic verbs and nouns that are based on three-letter roots, like the word "شعر" which can have different meanings depending on the different combinations of vowels ("to feel", "poetry", "hair", etc.) or the word "علم" which can mean "flag", "science", etc.

Another reason of this dissymmetry is related to the agglutination feature of Arabic words. The agglutination happens when a minimal form of a word is attached to various proclitics (interrogative style, similarity, link, etc.) or to various enclitics (mainly to add pronouns). These three examples can illustrate various situations of agglutination: (1) Kateb Yasseen كاتب يسين, (2) Does Kateb Yasseen ? أكاتب يسين, (3) I write to Yasseen أكاتب يسين.

2.2 Information retrieval in Arabic language

Most of the search queries, whatever the languages are, concern named entities such as proper names, etc. In another hand, to check the linguistic structures of the queries, we have made some tests using a sub-set containing 2850 Arabic queries that have been submitted to a multilingual Web directory described in (Boualem & al., 2001). These tests allowed us to see that 94,2% of the queries concern nominal structures, only 3,30% concern verbal structures and only 2,5% concern grammatical words. In fact these results can be lightly adjusted if we consider that queries do not contain vowels. Thus, apart some verbal structures and some non-ambiguous grammatical words such as من, متى, اخترع , most of the queries are very ambiguous (نزل, رقص, طلب , etc).

Concerning Arabic proper nouns, they often are derived from verbal structures (active participle, passive participle, etc.). For example كاتب means "author" and also is a part of a proper name such as in Kateb Yasseen. However, searching كاتب generally retrieves "author".

To argue these observations we present here some examples of search queries using Google :

- for the keyword كَتَب (with vowels such as "katab"), the first results are related to "books" (which transcription is "kutub") : in this case can we consider these results as a consequence of the "ranking" algorithm or is it related to a kind of "priority" for nouns ? Anyhow we can easily see that adding vowels to the keywords has no influence on the searching results.

- when searching for **جمال الدين الأفغاني** or **جمال عبد الناصر الزعيم**, the keyword جمال first retrieves adjectival results related to "beauty", and not related to the proper noun جمال. More precisely, we get 5 340 000 answers for **جمال**, 737 000 answers for **جمال الدين**, and 70 700 answers for **جمال الدين الأفغاني**. When searching for **ناصر** we get 805 000 answers for **جمال عبد**, 293 000 answers for **جمال عبد الناصر** and 253 000 answers for **جمال عبد الناصر الزعيم**. In the same way, the keyword الزعيم, retrievers 2 100 000 answers for **الزعيم**. In this case we noticed that the first displayed results are related to some soccer blogs, or related to theatre information about Adel Imam (742 000 answers for **الزعيم عادل**). Hence the first result related to **جمال عبد الناصر** comes in the 30th position. We conclude that there is a significant lack in processing named entities.

3 Benefits of natural language processing for information retrieval in Arabic language

3.1 Lemmatisation

We think that information retrieval is somehow language-dependent in the sense that search engines should adapt indexation and searching strategies to the language specificities. For Arabic, which has a complex (even regular) morphology (Dichy, 1990), we think that search engines should primary focus at least on lemmatisation. We try here, through some examples of some linguistic phenomena, to show the limitations of "artificial linguistic processing" in the indexation process and the benefits of lemmatisation for information retrieval in Arabic language.

Arabic has a very flexional morphology where morphological families can reach huge numbers of combinations. A lot of graphical forms of words, even they seem very similar, might not belong to the same semantic families and even to the same morphological families. Let us see for example the search results for the derived forms of the word قال : the query « *قال* » provides more than 146 forms, which largely exceeds the derivational combinations of this word. Indeed, the query retrieves words such us الانتقال , اعتقالهم , استقالة , العقال , مقاليد , اقالتهم , التقاليد , برتقالية , اثقالاً , وقالباً , الأقاليم , قالب , ... Moreover besides this huge "noise", a lot of other morphological variations of this word need to be found through other queries (e.g. imperfective يقول and other related deverbal forms).

For another example with the keyword « سماء », Google retrieves 594 000 answers by applying a completion method. Results also contain 279 000 answers for **أسماء**, where we can also find **الأسماء** (which is a rare plural form of سماء used for example in titles such as أسماء السماوات.). We think that these morphological and semantic distances are due to the fact of

applying to Arabic language the lemmatisation rules of other languages such as English.

In another hand, even applying Arabic lemmatisation rules does not allow obtaining good results in information retrieval because the derivational system of Arabic is more complex than just using suffixes. However lemmatisation in searching Arabic is necessary due to the agglutinant specificities of Arabic words completed by using proclitics and enclitics.

3.2 Gender, number and lemmatisation

To enrich our analysis about lemmatisation in searching Arabic we focus now on two nominal aspects, gender and number and we try to show the limitations of indexing techniques.

3.2.1 Singular and plural

Let us consider the plural word كتابات, the "standard" lemmatisation procedure, which consider making the plural form by adding the ات suffix to the singular form, should give the lemma كتاب , but the right lemma is كتابة. The same problem can be found when trying to lemmatise the dual form فتاتان to the singular form فتات, but the right one is فتاة. Our tests on Google have shown its limitations in processing theses kind of linguistic phenomena when confusing word terminations ت and ة. The "broken plural", which is a non-regular plural in Arabic language and that does not follow any flexional rules, comes to add more complexity to the lemmatisation procedures (رجال-رجل, for man-men and نساء-نسوة-امرأة for woman-women). Also some dual forms, in a morphological point of view, might correspond to singular forms, such as for example the country noun البحرين or the personal pronoun محمدين.

We also have analyzed the user's queries and have extracted the following information about using singular, dual and plural forms in keywords :

Number				
Singular	Dual	Plural		
74,21%	1,77%	24,02%		
		Regular masculine	Regular feminine	Broken plural
		71,09%	21,29%	7,52%

3.2.2 Masculine and feminine

Suffixation rules in general can be used to obtain masculine and feminine forms. To obtain a feminine form, the "standard" rule aims to add the suffix ة to the masculine form. However (again) this rule can not be always systemised, such as in the feminine words اثارة and دراسة which do not have masculine forms. Also there are many masculine Arabic word ended by the letter ة, such as in the word خليفة. In another hand, the gender might also be expressed through different words having different roots, like for example these masculine forms رجل أب ولد حصان جمل.

We also have analyzed the user's queries and have extracted the following information about using masculine and feminine in keywords :

Gender					
Masculine	Feminine				
50,13%	49,84%				
	With suffix ة		Without suffix ة		Others
	with masculine	without masculine	with masculine	without masculine	Feminine of masculine plural
	47,11%	11,69%	16,81%	1,01%	23,38%

4 Conclusion

Arabic information retrieval still faces many difficulties due to the Arabic linguistic features, especially its complex morphology and the absence of vowels in available documents and texts. These specificities often cause significant dissymmetry between the indexation process and the query analysis. We have presented in this paper some of the morphological constraints of Arabic language and we have shown through experimental tests how search engines deal with them. Finally this paper clearly states that information retrieval in Arabic language will never succeed without including language processing tools at all the linguistic levels (lexical, syntactic and semantic).

5 References

Abbes, R. (2004). La conception et la réalisation d'un concordancier pour l'arabe. Thèse de doctorat en Sciences de l'Information, Lyon : INSA, décembre 2004.

Abbes, R. & Dichy, J. (2008). Extraction automatique de fréquences lexicales en arabe et analyse d'un corpus journalistique avec le logiciel AraConc et la base de connaissances DIINAR.1, JADT'2008, 12-13 mars 2008, ENS-LSH Lyon, France.

Abbes, R. & al. (2004). The Architecture of a Standard Arabic Lexical database: some figures, ratios and categories from the DIINAR.1 source program. In : COLING'04, 20th International Conference on Computational Linguistics. Proceedings of the Workshop Computational Approaches to Arabic Script-bases Languages, 28.08.2004, Genève : 15-22.

Buckwalter, T. (2004). Issues in Arabic Orthography and Morphological Analysis. In : COLING'04, 20th International Conference on Computational Linguistics. Proceedings of the Workshop Computational Approaches to Arabic Script-bases Languages, 28.08.2004, Genève : 31-41.

Boualem, M. & Sneifer, R. (2001). Hahooa Arabic Web Directory & Natural Language Processing for Arabic Information Retrieval, ACL 2001, Workshop on Arabic Language Processing, Toulouse, France, July 2001.

Boualem, M. (1995). Arabic language processing, SNLP'95 Symposium on Natural Language Processing, Bangkok, Thailande, Août 1995.

Boualem, M. & Zajac, R. (1999). Unicode-based Arabic text, ATLAS'99, Arabic Translation and Localisation Symposium, Tunis, May 26-28, 1999.

Brin, S. & Page, L. (1988). The anatomy of a large-scale hypertextual web search engine, Computer Networks and ISDN Systems 30 (1988), pp. 107–117.

Chen, P. & al. (2007). Finding scientific gems with Google's PageRank algorithm, Journal of InformetricsVolume 1, Issue 1, January 2007, pp. 8-15.

Dichy, J. (1990). L'écriture dans la représentation de la langue : la lettre et le mot en arabe. Thèse d'État, Université Lumière-Lyon 2.

Harman, D. (1991). How effective is suffixing? Journal of the American Society of Information Science. vol. 42, No 1. pp. 7-15, 1991.