Addaall Arabic Search Engine: Improving Search based on Combination of Morphological Analysis and Generation Considering Semantic Patterns

Mamoun Hattab[*], Bassam Haddad[†], Mustafa Yaseen[‡], Asem Duraidi[*] and A. Abu Shmais[*]

*Arabic Textware Inc., Amman-Jordan
{m.hattab, a.duraidi, a.abushmais}@arabtext.ws
†University of Petra, Department of Computer Science, Amman-Jordan
haddad@uop.edu.jo
‡Amman University, Department of Computer Science, Amman-Jordan
myaseen@ammanu.edu.jo edu.jo

Abstract

This paper addresses some issues involved in utilization of different levels of Arabic morphological knowledge in improving the search process in a search engine. The different levels of morphological knowledge can be considered as an incremental process considering the next sensitive word pattern in the context of enhancing the possibility of covering a higher recall and precision.

1. Introduction

This paper attempts to demonstrate how the utilization of different levels of Arabic morphological knowledge can improve the searching process in a Search Engine. We can show how combining *morphological analysis* and *generation* can be used for retrieving information based on *semi-semantic linguistic* search in Arabic. This technique extends the usage of roots and stems into a method of categorization for morphological patterns into semantic groups, and uses a morphological generator to produce the different inflections of a word to be retrieved based on the semantic distance from the word.

Google approach is primary based on the matching process between the user's key words and the texts under which it is indexed in terms of words (Brin S. and Page L., 1998). Words are treated as a set of symbols, not words with meanings to human users. And while plain linguistic measures, such as stemming and structured data search are used to enhance results, in nature Google and such search engines are still "Symbolic Computing" machines.

In third generation of search engines, Natural Language Processing technologies are applied in searching extensively, because in the first place, the search is seen as a language understanding process. This approach for search is paradigmatically different and a level higher in terms of the degrees of system difficulty and complexity. And it offers more accurate and consistent search results than the second generation search engines through its intelligence in language understanding. In this context many researchers have considered this aspect (Abuleil and Alsamra, 2004; Chen and Gey, 2002; Aljlayl M., 2002).

For a language like Arabic where the structure of a word can change according to many factors while maintaining the same meaning, *"Symbolic Computing"* does not retrieve an accurate set of results and there is an immense need for a linguistic approach to handle the search process (Brin S. and Page L., 1998).

Addaall[1] Arabic Search Engine built by Arabic Textware Inc[2]. utilizes a morphological analyzer and generator to construct different indices based on both the root and stem of a word. While retrieving information based on the root overflows the user with a complete but less relevant set of results, the use of the stem based search not only retrieves results in lesser numbers but also the set of results is less accurate.

2. Morphological Knowledge implies Syntactic and Semantic Indicators

There are much information produced by the morphological analysis which are useful in the next steps of processing such as the syntactic and the semantic analysis. Some of the information is:

- **The morphological type of a word**

  We mean by *"morphological type"* determining if the word is noun, verb, or particle; and what subtype of nouns, verbs, particles it is. It is clearly noticed that the syntactic functionality of the word is highly dependent on the morphological type of the word. For example, the syntactic function of a *"subject"* requires that the morphological type is a noun. The morphological type provides the semantic analysis with the semantic properties of each word. For example, the determining of (/فَعّال/, fa‘‘ āl) to be an exaggeration form tells that the meaning is much

[1] http://www.addaall.com
[2] http://www.arabtext.ws/

emphasized.

- **The Determiner**

  We mean by "determiner" any indeclinable noun, verb, determiner or particle in addition to the affixes. It is clear that the variety of the syntactic functions which the determiners can do is wide. The proposition (/فِي /, fī, in) provides that the word follows is a noun and genitive (Haddad B., 2007). Even the letters such as (/و /, wāw) in (/مُرسلو /) which is a pronoun, tells that the whole word is sound masculine plural, which affects the whole sentence.

  One of the examples of how the determiners benefits the semantic analysis is the prefix (/ال /l)[3] which tells that the word following is a definite noun. Furthermore the state of indefinite and dual is which can extracted from the morphology can also semantically be considered as a source if semantic knowledge such as unique quantification. Another example is the prefix (/ س /, s) that direct the meaning of present verb to future.

- **The Morphological Pattern of a word**

  It is true, even not clear for every one, that the patterns of Arabic words include syntactic and semantic content. The pattern (/فَعُلَ/) of the past verb tells that the verb is intransitive which means that the sentence is complete with no need to have an object. An example of the semantic profit is the meaning of request that given by the pattern (/استفعل/) and its derivatives.

- **The root**

  The root: All roots have a semantic aspect because the root in Arabic is the part which contains the meaning; it is the core of language. Anyway, some roots has syntactic properties like (/حسب /) which tells that the verb needs two objects.

2.1 Morphological Levels for Search

To measure the closeness between the meaning of the word being searched for, and the suggested words to be searched for, we adopted the concept of utilizing different levels of morphological knowledge. The least degree of relationship is the strongest between the original word and the alternatives.

In the Zero Level the same and identical word is considered in the search process.

Level One:

A. For the verbs: verb inflections referring to the same tense. (e.g. ضربتم /ضربوا /ضربَ)

B. For the derivative nouns: its grammatical states (قائلٌ، قائلاً، قائلٍ، القائلُ، القائلَ، القائلِ).

C. For the gerund: its grammatical states (قولٌ، قوْلاً، قوْلٍ).

Level Two:

A. For the verb: verb inflections referring to all tenses. (e.g. تضربان/ اضرب/ يضرب/ ضربَ...)

B. For the Derivative noun: its grammatical and morphological states (قائلٌ/ قائلةٌ/ قائلان/ قائلين، قائلاتٌ، قائلات، القائلون، القائلاتُ، القائلاتَ.).

C. For the gerund: its grammatical and morphological states (قولٌ، قوْلاً، قوْلٍ، القوْلُ، القوْلَ، القوْلِ).

Level Three:

A. Connecting the verb only to its gerund and vice versa (e.g. الضربِ/ ضرْب/ ضرَبَ).

B. Connecting the derivative noun to its corresponding morphological counterparts that share the same morphological category. For Example connecting exaggeration pattern together such as (قوّالٌ/ / مِقوال ...)

Level Four:

A. Connecting the verb to its gerund and its derivative nouns and vice versa (e.g. مستقيل /استقال/ مستقال /).

B. Connecting the gerunds and the derivative nouns to each other (استفادة/ مستفاد/ مستفيد).

[3] Please notice that deep semantic analysis considering the logical compositionality of determiners such as (/ال /, ‹l) is out of scope of this paper. Determiners in the logical sense (Haddad B. 2007) represent quantifiers and a special case of Generalized Arabic Quantifier (GAQ) such as:

$$\begin{bmatrix} (/\text{معظم-ال}/, \text{most-of-the}) \\ \text{CAT} \quad \text{DET}_{\text{Arab}} \end{bmatrix}_{\text{sem}} \equiv \lambda R.\lambda S.(\ /\text{معظم-ال}(x)/, \text{most-of-the}(x)\ ).\ (R, S), \text{ where } |R \cap S| > |R - S|$$

, whereas and in this context (/ال /, ‹l) can be regarded as a numeric quantifier:

$$\begin{bmatrix} (/\text{ال}/, \text{The}) \\ \text{CAT} \quad \text{DET} \\ \text{AG} \quad [\text{NUM} \quad sing] \end{bmatrix}_{sem} \text{ as } \left[(/\text{ال}_{\text{I}}\ /, \text{The}_1)\right]_{sem} \text{ where}$$

$$\left[(/\text{ال}_{\text{I}}\ /, \text{The}_1)\right]_{sem} \equiv \lambda P.\ \lambda Q\ .\exists x.(\forall y\,(P(y) \Leftrightarrow x = y) \wedge Q(x))$$

And finally there is the Root **Level,** representing the most possible search in the context of semantic dependency.

3. Enhancing Search based on different level of Morphological Knowledge

To enhance the quality of results by minimizing the number of results while maintaining more accurate relevancy and precision, a hybrid approach combining morphological analysis and generation was applied. In this approach a query goes through the following processes:

- Morpho-Syntactic analysis to define its root/stem and part of speech.
- Morphological generation to produce the nearest inflections to that word based on a semantic categorization of the different patterns that may apply to that root/stem.

When you search based on a word's root, you are sure to get all of the relevant results, but with other irrelevant ones. Because although each result retrieved includes at least one inflection of the word, this inflection might not be necessarily relevant.

This feature of Arabic language can not be controlled without having a lexicon that stores semantic knowledge. In that case, all the words will be defined with their semantic relationships to other words in the lexicon, extending the possibilities to cover not only inflections but synonyms and even related conceptual and ontological knowledge.

In the case of stem search, stemming techniques for Arabic are not very well defined until now. The definition of a stem in Arabic is often exchanged with that of the root. Many researchers even argue that if the stem has a different meaning than a root then it does not apply to Arabic language. They consider it a foreign concept.

When applying stemming to an Arabic word, we mean that we eliminate or remove prefixes and suffixes from that word. What is left is a partial word that in many cases has possibly no meaning and does not exist in the dictionary as it is. We then use this partial word to do partial search looking for words that include this part.

The outcome of such a search has fewer results than the search by root. But how significant are these in term of precision and recall? That is why we believe that Arabic stem search is neither accurate nor comprehensive.

This led us to suggest a method to improve Arabic search techniques by using the morphological generator. In order to do so, we classified the morphological patterns of Arabic semantically according to their meanings. For each word we want to search for, the most morphologically and semantically related word form will be generated to be a subject of the improved search process.

An Example:

If the user is searching for (/ضَرَبَ /, hits) then we notice the following:

In the same level of (/ضَرَبَ /, hits) is the only suggested alternative.

In next level or the first one, the suggested alternatives are (ضَرَبَ/ ضَرَبوا/ ضَرَبْتم/ ضَرَبَا/ ضَرَبْتَ... etc.)

The next level would suggest alternatives such as (ضَرَبَ/ ضَرْب/ الضربِ...etc.)

In next level the suggested candidates are (يضرب/ ضربَ/ اضرب/ تضربان...etc.)

In fourth level the suggested alternatives are (ضارب/ ضَرَبَ/ مضروب/ مضروبتين...etc.)

In last level the suggested alternatives are all the nouns and verbs that have the root (ضَرَبَ).

4. Conclusion and Outlook

In this paper, we have tried to stress on the importance of utilizing different level of morphological knowledge towards improving the quality and quantity of the Addaall Arabic Search engine. The results are very promising as the coverage and the sensitivity of this approach is relatively incredibly practical (see link to the Search Engine http://www.addaall.com).

Furthermore, our approach has considered the concept of categorizing Arabic patterns according to their meanings, whereas some *semi-semantic information* has been useful in the search process. In this context, we have established a matrix correlating each pattern of speech to semantically related pattern. The different levels of morphological knowledge can be considered as an incremental process considering the next sensitive word pattern; i.e. to increase the recall, and as a predication and heuristic process to increase the precision; as heuristically related word might occurred in a near distance.

Finally, although we understand that this approach produces a relative and not complete set of results, we can see that it in fact represents a significant improvement towards Arabic search engines in view of its practical and pragmatic use in real implementations.

References

Abuleil S., Alsamra K. (2004). New Technique to Support Arabic Noun Morphology: Arabic Noun Classifier System (ANCS). International Journal of Computer Processing of Oriental Languages, Vol. 17 Issue 2.

Aitao Chen and Fredric Gey (2002). Building an Arabic stemmer for information retrieval. The Eleventh Text Retrieval Conference (TREC 2002), National Institute of Standards and Technology (NIST).

Aljlayl M. (2002). Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach. ACM Eleventh Conference on Information and Knowledge Management.

Brin S. and Page L. (1998). The anatomy of a large hypertextual web search engine. Web publication, Stanford University.

Haddad Bassam (2007). Semantic Representation of Arabic: A Logical Approach towards Compositionality and Generalized Arabic Quantifiers. International Journal of Computer Processing of Oriental Languages, Vol. 20. Nr. 1. 2007