Effective Stemming for Arabic Information Retrieval

Youssef Kadri & Jian-Yun Nie
Laboratoire RALI, DIRO, Université de Montréal
{kadriyou,nie}@iro.umontreal.ca

Arabic has a very rich and complex morphology. Its appropriate morphological processing is very important for Information Retrieval (IR). In this paper, we propose a new stemming technique that tries to determine the stem of a word representing the semantic core of this word according to Arabic morphology. This method is compared to a commonly used light stemming technique which truncates a word by simple rules. Our tests on TREC collections show that the new stemming technique is more effective than the light stemming.

Keywords: Arabic morphology, linguistic-based stemming, light stemming.

1. INTRODUCTION

Arabic language raises several challenges to Natural Language Processing (NLP) largely due to its rich morphology. In this language, morphological processing becomes particularly important for Information Retrieval (IR), because IR needs to determine an appropriate form of words as index. Arabic word stemming has been a central topic of many researches in Arabic IR. Khoja (Khoja, 1999) attempts to find roots for Arabic words which are far more abstract than stems. It first removes prefixes and suffixes, then attempts to find the root for the stripped form. McNamee (McNamee, 2002) uses the matching n-grams of multiple lengths to index words which generates a big size index. Light stemmers developed by Larkey (Larkey, 2001), Darwish (Darwish, 2002) and Chen (Chen, 2002) select some prefixes and suffixes to truncate from the words. This last approach is inspired by the stemming process of English. Because it gives the best performance, this approach is widely used now in IR.

Despite these studies, it is still unclear what type of stemming is appropriate for Arabic IR. On one hand, a light stemming may prevent from grouping two different words; but it also runs the risk of failing to group two semantically similar words, leading to a low recall. On the other hand, a too severe stemming may incorrectly group semantically non similar words into the same index, leading to a low precision. More investigations on the effects of stemming on IR effectiveness are needed.

In this study, we propose a new stemming method that tries to determine the core of a word. Our method is motivated by the composition of words in Arabic: Arabic words are usually formed as a sequence of {antefix, prefix, core, suffix, postfix}. We believe that a good stemming strategy is to determine such cores as indexes. In so doing, the indexes would encode the basic semantics in Arabic language.

The above method is compared to a light stemming technique which truncates a word at the two ends. This method is similar to that proposed by Larkey, Darwish and Chen. We tested the two stemming methods on TREC collections and the results show that the new stemming technique is more effective than the light stemming.

The remainder of this paper is organized as follows: We will first describe the basic characteristics of Arabic morphology and the problems related to its processing in section 2. Section 3 describes the two stemming techniques. Section 4 is devoted to the description of the experiments. Finally, we present the results of our experiments with analysis.

2. ARABIC MORPHOLOGY

Arabic has an origin very different from European languages. It includes 28 letters and it is written cursively from right to left. The morphological representation of Arabic is rather complex because of the morphological variation and the agglutination phenomenon (Kadri, 1992). Letters change forms according to their position in the word (beginning, middle, end and separate). Table 1 gives an example of different forms of the letter "gh" at different positions. We can observe several general characteristics of this language as follows:

Beginning	Middle	End	Separate
غ	غ	غ	غ

TABLE 1: Different writings of the letter "*gh"* at different positions within word or as a separate letter

- Most of nouns and verbs are derived from a reduced number (approximately 10 000) of roots. These roots are linguistic units carrying a semantic meaning and most of these roots consist of only 3 consonants, rarely 4 or 5 consonants.

- From these roots, we can generate nominal and verbal derivatives by the application of the templates (morphological rules). One can generate up to 30 words from a 3 consonants root. Table 2 shows an example with the 3-grams root "*ktb"* (to write), from which we can produce several words:

Write	Book	Writer	Written	Small book
كتب	كتاب	كاتب	مكتوب	كتيب

TABLE 2: Derivation of several words from the "*ktb"* root

- In written Arabic, the vowels (diacritics) are omitted and as a result of this omission, the words tend to have a higher level of ambiguity (Kadri, 1992). For example, the word (على) without vowels can mean the proper name (Ali) or the preposition (on). This ambiguity will be a crucial problem in information retrieval in the fact that an Arabic word can have several meanings.

- In addition to the ambiguity phenomenon, there is another problem of the plural form of irregular nouns, also called broken plural. In this case, a noun in plural takes another morphological form different from its initial form in singular. The absence of a dictionary for these irregular nouns makes it difficult to design a rule-based algorithm to transform this kind of plural to singular form.

- Words are separated by space and other punctuation marks. Nevertheless, prepositions are agglutinated to the word appearing after them, making the boundary between the word and the preposition invisible.

- Several types of affix are agglutinated to the beginning and the end of the words: antefixes, prefixes, suffixes and postfixes. One can categorize them according to their syntactic role. Antefixes are generally prepositions agglutinated to words at the beginning. Prefixes, usually represented by only one letter, indicate the conjugation person of verbs in the present tense. Suffixes are the conjugation terminations of verbs and the dual/plural/female marks for the nouns. Finally, postfixes represent pronouns attached at the end of the words. All these affixes should be treated correctly during word stemming.

Obviously, Arabic is also very different from the European languages at syntactic and semantic levels. However, this is beyond the focus of this paper.

3. STEMMING

The objective of stemming is to find the representative indexing form of a word by the application of truncation of affixes. As we stated, there are four kinds of affixes: antefixes, prefixes, suffixes and postfixes that can be attached to words. Thus an Arabic word can have a more complicated form if all these affixes are attached to its root. The following table shows an example of word (**ليفاوضونهم**) with all types of affix:

Antefix	Prefix	Core	Suffix	Postfix
لـ	يـ	فاوض	ون	هم
Preposition meaning “to”	A letter meaning the tense and the person of conjugation	negotiate	Termination of conjugation	A pronoun meaning “them”

TABLE 3: An agglutinated form of an Arabic word meaning “to negotiate with them”

For this example, it will not be enough for IR if we truncate only one prefix and only one suffix from this word. The resulting form (**يفاوضون**) will not be common to other semantically similar words. For example, a very similar word (**ليفاوضهم**) is stemmed with (**يفاوض**). We see clearly that even if the two words are semantically similar, their stems are different. We believe that if we arrive to truncate all these affixes from the first word, we will obtain a better form of index. This form represents the semantic core of the word. Therefore, we propose a linguistic-based stemming method that tries to determine the core of a word.

3.1 Linguistic-based stemming

This method is motivated by the composition of words in Arabic: Arabic words are usually formed as a sequence of {antefix, prefix, core, suffix, postfix}. The following table describes the affixes used in Arabic:

Antefixes	Prefixes	Suffixes	Postfixes
وبال, وال, بال, فال, كال, ولل, ال, وب, ول, لل, فس, فب, فل, وس, ك, ف, و, ب, ل	ا, ن, ي, ت	تما, يون, تين, تان, ات, ان, ون, ين, وا, تا, تم, تن, نا, ت, ن, ا, ي, و	كما, هما, كن, هن, تي, ها, نا, هم, كم, ك, ه, ي
Prepositions meaning respectively: and with the, and the, with the, then the, as the, and to (for) the, the, and with, and to (for), then will, then with, then to (for), and will, as, then, and, with, to (for)	Letters meaning the conjugation person of verbs in the present tense	Terminations of conjugation for verbs and dual/plural/female marks for nouns	Pronouns meaning respectively: your, their, your, their, my, her, our, their, your, your, his, my

TABLE 4: Arabic affixes

A straightforward method would be to truncate possible affixes according to the above table. It does not go any further than finding the 3-gram root from the stripped form as Khoja attempts to do. However, we encounter many cases of ambiguity: a particular sequence of letters may or may not play a role of affix, depending on the word. No morphological rules are currently available to allow us to determine the correct affixes. Therefore, we take advantage of corpus [1] statistics: we apply rules that generate a set of possible stems for a word; then we select the most appropriate candidate according to corpus statistics - the most commonly used stem is selected. Corpus statistics are compiled on the 523 359 different words in the TREC collection. Each word of the collection undergoes different decompositions to obtain all possible stems for this word. By doing so for all words, we construct a corpus of stems along with their occurrence frequencies in the collection. Notice that this approach is reasonable because the stem is also a word that appears in the texts. Corpus statistics can thus reveal the most commonly used stems in Arabic. The selection of this common form of stem can solve most of the ambiguities.

3.2 Light stemming

This approach is statistically motivated. It is similar to the commonly used light stemmers. It truncates a word at the two ends. The decision to truncate or not a segment of a word is made according to some rules and statistics on the corpus (Kadri, 2004). We grouped all affixes in 2 classes: prefixes and suffixes. Then we made a statistics table based on the occurrence frequencies of these affixes on the 523 359 different tokens on the TREC collection [1]. Finally, we set a list of the most frequent prefixes to remove from the beginning of words. These prefixes (وبال, فل, ولل, كال, فال, ول, وب, بال, لل, والل, ال, ب, ل, وال, و, ا , فب) are generally prepositions or sometimes several prepositions attached to the words. Suffixes that we judged necessary to truncate from words are those which are the most frequent and represent generally pronouns expressing number or gender of Arabic nouns: (ا, ه, ن, ي, ت, ها, ين, و, ان, ات, ون, هم, نا, ك, وا, هما, تي). We notice also that our method shares several prefixes and suffixes to be removed with the light stemmers developed by Larkey, Darwish and Chen.

4. EXPERIMENTS

The goal of our experiments is to compare the two stemming methods for IR. We used the Arabic TREC collection [1] which contains 383 872 documents, selected from AFP (France Press Agency) Arabic Newswire. These documents are newspaper articles covering the period from May 1994 until December 2000. We use two sets of topics: TREC 2001 containing 25 topics and TREC 2002 containing 50 topics. We used the title and the description fields of topics.

4.1 Morphologic pre-processing

In written Arabic, diacritics are often omitted in texts and a familiar reader with this language will not have difficulties to read correctly a text without vowels. In addition, the letters change forms according to their position in the word (beginning, middle, end and separate). Some of these letters undergo a light modification in writing which does not influence considerably the meaning of the word. For example the letter "ا" at the beginning of a word can take different forms: "إ" ,"أ" or "آ". Regarding all these specificities of this language and in order to overcome the problem of the representation variation of Arabic letters, we applied some normalization methods on both the documents and the topics before indexing:

- Replacing "إ" ,"أ" and "آ" by alif bar "ا".
- Replacing "ى" by "ي" at the end of the words.
- Replacing "ة" by "ه" at the end of the words.
- Replacing the sequence "يء" by "ئ".
- Removing the tatweel character "-", used for aesthetic writing in the Arabic texts.
- Removing the shedda "ّ" and the diacritics.

4.2 Stopwords

As in other languages, Arabic also contains functional words (or stop words) which do not carry a particular and useful meaning for IR. Thus, we set up a stop list, which contains 413 function words: prepositions, particles of Arabic, and the translation of some English stop words.

4.3 The retrieval model

The retrieval model is a unigram language modeling algorithm based on Kullback-Leibler divergence [12]. Given a query Q and a document D, we compute the relevance score of this document to the query with the negative of the divergence of the query's language model from the document's language model (Zhai, 2001-a):

$$R(Q,D) \propto \sum_t p(t \mid Q) \log p(t \mid D) \qquad (1)$$

To avoid the problem of attributing zero probability to query terms not occurring in document D, smoothing techniques are used to estimate $p(t|D)$. One uses the Jelinek-Mercer smoothing technique which is a method of interpolating between the document

and collection language model (Zhai, 2001-b). The smoothed *p(t|D)* is calculated as follows:

$$p(t \mid D) = (1-\lambda)\frac{tf(t,D)}{|D|} + \lambda p(t \mid C) \qquad (2)$$

where $\frac{tf(t,D)}{|D|}$ and $p(t \mid C)$ are the maximum likelihood estimates of a unigram language model based respectively on the given document *D* and the collection of documents *C*. λ is a parameter that controls the influence of each model.

4.4 Experimental results

Before indexing, Arabic characters are normalized on both the documents and the topics collections (See section 4.1). Stop words are removed and documents are ranked to each topic according the formula (1) in section 4.3. In our experiments, we use the classical IR measure of Mean Average Precision (MAP) on the eleven recall points. Recall is also used as a second measure. The table below presents the results of all these experiments:

Query collection	Measures	Linguistic-based stemming	Light stemming
TREC 2001 (25 topics)	MAP	0.3326	0.3220
	Recall / 4122	2704	2664
TREC 2002 (50 topics)	MAP	0.2828	0.2671
	Recall / 5909	4301	4443
Merged TREC 2001-2002 (75 topics)	MAP	0.3107	0.2868
	Recall / 10 031	7181	7121

TABLE 5: Arabic monolingual IR performances according the two stemming methods

On both sets of queries, the results show that the new stemming technique we propose results in consistently better retrieval effectiveness than the light stemming technique. We obtained 31 % of MAP with the linguistic-based stemming method on the merged topics collection against 28 % for the light stemming technique. This result shows that a light stemming in Arabic language is not the best approach for Arabic IR: it fails to group many semantically similar words into the same index. In contrast, our method can better determine the semantic core of a word. The most the indexes group semantically similar words; the most the performance of IR is better. We give some examples to make comparison with the two stemming methods:

Words	Linguistic-based stemming	Light stemming
عراقيين	عراق	عراقي
البوسنيه	بوسن	بوسني
مهرجان	مهرجان	مهرج

TABLE 6: Stemming results of some words according to the two methods

The above table shows the results of stemming of three words according to the two stemming techniques. The linguistic-based stemming method produces better stems than the light stemming. For the two first examples (**عراقيين**, **البوسنيه**), the new method extracts all affixes and determines stems correctly. In contrast, light stemming fails to

extract all affixes. It fails to determine a one-letter suffix (ي) and consequently produces a stem that is not representative for many semantically similar words. For the third example (مهرجان), we take advantage of corpus statistics to solve a case of ambiguity. The two letters (ان) do not represent a suffix for this word. Even though it is not a suffix, light stemming truncates it from the word and produces an erroneous stem (مهرج). In contrast, our method, which uses corpus statistics, applies different decompositions, results in a set of candidate stems and selects the most commonly used stem in the corpus.

5. CONCLUSION

In this study, we looked at the problem of stemming Arabic words for the purpose of IR. IR needs to determine an appropriate form of words as index. We propose a new stemming method that tries to determine the core of a word according to linguistic rules. This method is compared to a light stemming technique. The new method shows better retrieval effectiveness than the light stemming. The light stemming fails to group many semantically similar words into the same index. In contrast, the linguistic-based method can better determine the semantic core of a word. However, the new method can also lead to some errors, since an affix which normally forms part of a word, may be truncated. To overcome this kind of ambiguity, we believe that this method can be improved at the level of corpus statistics. More processing must be done to select the correct stem when different candidate stems are possible for a word.

ENDNOTES

[1] Arabic TREC collection. It contains 523 359 different words. http://trec.nist.gov/

REFERENCES

Khoja S. and Garside S. (1999). 'Stemming Arabic Text'. Computing Department, Lancaster University, Lancaster, U.K. http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps, September 22, 1999.

McNamee P., Piatko C. and Mayfield J. (2002). 'JHU/APL at TREC 2002: Experiments in Filtering and Arabic Retrieval'. TREC-11 conference 2002.

Larkey L. S. and Connell M. E. (2001). 'Arabic information retrieval at UMass in TREC-10'. TREC-10 conference, Gaithersburg, Maryland 2001.

Darwish K. and Oard D. W. (2002). 'CLIR experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval'. TREC-11 conference 2002.

Chen A. and Gey F. (2002). 'Building an Arabic stemmer for information retrieval'. TREC-11 conference 2002.

Kadri Y. and Benyamina A. (1992). 'A syntax semantic analyzer for Arabic language'. Engineer thesis, University of Oran 1992.

Kadri Y. (2004). 'Query translation for English-Arabic cross language information retrieval'. TALN conference 2004.

Zhai C. and Lafferty J. (2001-a). 'Model-based feedback in the language modeling approach to information retrieval'. Tenth International Conference on Information and Knowledge Management (CIKM), 2001.

Zhai C. and Lafferty J. (2001-b). 'A study of smoothing methods for language models applied to ad hoc information retrieval'. Proceedings of the ACM-SIGIR, 2001.