Journal of Software Engineering and Applications, 2011, 4, 522-526
doi:10.4236/jsea.2011.49060 Published Online September 2011 (http://www.SciRP.org/journal/jsea)

The Enhancement of Arabic Stemming by Using Light Stemming and Dictionary-Based Stemming

Yasir Alhanini, Mohd Juzaiddin Ab Aziz

School of Computer Science, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia.
Email: din@ftsm.ukm.my

Received August 9th, 2011; revised September 2nd, 2011; accepted September 9th, 2011.

ABSTRACT

Word stemming is one of the most important factors that affect the performance of many natural language processing applications such as part of speech tagging, syntactic parsing, machine translation system and information retrieval systems. Computational stemming is an urgent problem for Arabic Natural Language Processing, because Arabic is a highly inflected language. The existing stemmers have ignored the handling of multi-word expressions and identification of Arabic names. We used the enhanced stemming for extracting the stem of Arabic words that is based on light stemming and dictionary-based stemming approach. The enhanced stemmer includes the handling of multiword expressions and the named entity recognition. We have used Arabic corpus that consists of ten documents in order to evaluate the enhanced stemmer. We reported the accuracy values for the enhanced stemmer, light stemmer, and dictionary-based stemmer in each document. The results obtain shows that the average of accuracy in enhanced stemmer on the corpus is 96.29%. The experimental results showed that the enhanced stemmer is better than the light stemmer and dictionary-based stemmer that achieved highest accuracy values.

Keywords: *Dictionary-Based Stemmer*, *Arabic Morphological Analyzer*, *Named Entity Recognition*

1. Introduction

Word stemming is one of the most important factors that affect the performance of many natural language processing applications such as part of speech tagging, syntactic parsing, machine translation system and information retrieval systems. In Arabic, there are two main approaches for stemming: light stemming and dictionary-based stemming. The light stemming is the affix removal approach that refers to a process of stripping off a small set of prefixes and/or suffixes to find the root of the word. There are some recent works that used the light stemming to extract the root or stem of Arabic words [1-5]. The main disadvantage of these works is that they ignore the identification of Arabic names that increase the ambiguity rate of the stemmer. Although light stemming can correctly generate the root or stem for many variants of words, but it fails to find the root of many words. For example, the broken (irregular) plurals for nouns do not get conflated with their singular forms, and past tense verbs do not get conflated with their present tense forms. On the other hand, the dictionary-based stemming is the morphological approach that depends on set of lexicons of Arabic stems, prefixes, and suffixes to extract the stem of words. This stemming can find the stem of the broken (irregular) plurals for nouns and irregular verbs, because the stem of these irregular words had been entered. Many researchers shed the light on the dictionary-based stemming to find the stem of Arabic words [6-9]. Although works have advantages and disadvantages, the main problem of this stemming is that it cannot deal with the words that are not found in the lexicon of stems. The dictionary-based stemming has the ambiguity in which it may give more than stems for the same word. The multi-word expressions are more complicated expression which undergoes inflections and lexical variation when words are being understood compositionally; their meaning is lost and adds to ambiguity problem, as component may be separately ambiguous. However, the most of existing works [8-12] do not handle the multiword expression before extracting the stem of the words. The handling of the multiword expressions is to avoid the needless analysis of structure, and to reduce the stemming ambiguity and time of stemming.

2. Related Work

For Arabic word stemming, there are two main methodologies: the dictionary-based stemming and the light stemming. Dictionary-based stemmers match every word with a word on a proper digitalized dictionary, correspond each word to its stem. For example, [4,13,14], proposed three strategies for Arabic language morphologies development which depend on the level of analysis. Firstly, it involves the analysis of Arabic at the level of the stem, and the use of a regular concatenation. Stem is the form least remarkably in one word, that is, without a word uninflected, suffixes proclitics, prefixes or enclitics. Arabic, and this is usually perfective, person, singular verb, in the case of nouns and adjectives are in the form of the singular indefinite. Secondly, analyzed Arabic words consist of roots, pattern as well as concatenations. A root is a series of three also seldom two or four characters that are called root, pattern and template of vowels or a combination of consonants and vowels with slots and the inclusion of radicals from the root. Thirdly, analyzed Arabic words also consist of root, template and vocalization, in addition concatenations. Reference [8] has developed broad-coverage lexical resource to improve the accuracy of their morphological analyzer. It was constructed by analyzing 23 established Arabic language dictionaries. This morphological analyzer references on a detailed lists of affixes, clitics and patterns, which were extracted from authoritative Arabic grammar books and were then cross-checked by analyzing words of three corpora: the Qur'an, the Corpus of Contemporary Arabic, and Penn Arabic Treebank and Sawalha and Atwell lexicon base. The morphological analyzer uses novel algorithms that generate the correct pattern of the words, deal with the orthographic issues of the Arabic language and other word derivation issues, such as the elimination or substitution of root letters, tokenize the word into proclitics, prefixes, stem or root, suffixes and enclitics, generate all possible vowelizations of the processed word, and assign morphological features tags for the word's morphemes. A light stemmer is not dictionary dependent, for that reason it is not able to use a criterion that an affix can be removed only if what remains is an existing Arabic word. For example, [13] proposed to extracted trilateral Arabic roots by provided an effective way to removed suffix and prefix from the inflected words. After that, match letters of roots to removed each infixes in the patterns. In this algorithm, followed many steps such as normalized corpus by remove stops words and punctuation also it mach with the patterns. Although this algorithm resolved many problems, however some words cannot used same rules when remove *Fa* ف or *waw* و from single prefix because it be original letter. For example, فارس, واحد and ورد. The accuracy from the corpus about 92% by used 10582 words from 72 abstracts. Furthermore, [12] performed equivalently similar khoja stemmer without root dictionary by used Arabic Trec-2001 collection. Taghva criticized Khoja stemmer firstly; root of the dictionary requires maintenance to ensure that words are newly discovered stem correctly. An addition, replaced a weak letters such as أ ي و with و sometimes produce a root which not related with original words by removed part of root. Instance word منظمات (containers) is stemmed to ظمأ (Thirsty) instead of نظم. For this algorithm defied set of diacritical mark that removed by stemmer and defined some sets of patterns. They employed four stemmers Arabic TREC collection Composed of 383,872 news stories to compared three approaches khoja, ISRI and light, with not stemming. The find ISRI, khoja and light stemmers were much better than no stemming. Also light stemmer has been a higher precision for the higher ranked documents. The precision for light stemmer on the shorter title queries was 0.480; description was 0.424 and narrative 0.282. Reference [2] proposed a new stemming algorithm which depend on Arabic morphology and creation lemmatizer in linguistics by assumed which lemmatization will be more Efficiency in tokenizing Arabic document which stemming by overcoming the stemming errors and reduced stemming cost by reducing unnecessary.

3. Materials and Method

This section describes the enhancement of Arabic morphological analyzer that uses the light stemming and dictionary-based stemming. In the enhancement make use of hybrid method so that the light stemmer is applied to identify the stem of the word without using Arabic stems or roots. Despite the fact that light stemmer is efficient in most cases, but it cannot deal with the irregular word in Arabic. In addition, it gives the wrong stems in some words. After applying the light stemmer for the word, the verification is applied in order to check whether the identified stem is the real stem or not. There are two probabilities of output of light stemmer. The first probability, the stem is null (the light stemmer cannot identify the stem of the word). The second probability, the stem is not null, but it has less than three original letters. The third probability is the stem is not null and have more than three original letters. For the first probability, there is no more process. For others probabilities, the dictionary-based stemmer is applied to extract the stem of the word.

3.1. Light Stemming

The steps of light stemming: common words identification, word segmentation, and matching the patterns. The first step is to identify the common words (حروف الجر, حروف النصب, أدوات النداء,) and the non-derivational nouns (علم, اسم جنس). This step is very important task in stemmer in order to reduce the stemming time. The second step is the word segmentation. Arabic word is composed of stem of word and affix that indicate the tense, gender, number. Also the clitics are attached to the word. Some clitics are attached to beginning of the word (prepositions and conjunctions) while others, such as, pronouns at the end. This stage is to segment the word into its components (affix, stem, and clitics) according to Arabic rules. The main formula of Arabic words is defined as the following:

clitics + prefix + stem + suffix + clitics

where, the clitics, prefix, and suffix are attached to word optionally. The final step (matching the patterns) is to extract the stem and root of the word by matching the word without its affixes to the Arabic patterns. This step is applied to extract the stem and the root of word as follows. Let *Len* is the length of word after removing the prefixes and suffixes, the patterns that their lengths equal to *Len* have been selected. For each selected pattern, the stem will be matched with the particular pattern to compute the similarity between them. The pattern that its similarity with the stem equals to *Len*-3 will be selected as the form of this stem. For example, the segmentations of the word "بالواقعة" are the "بال" as prefix, and the "واقع" as a stem, and the "ة" as suffix. The stem "واقع"will select the pattern"فاعل" and the root of this word is "وقع".

3.2. Dictionary-Based Stemming

The dictionary-based stemming is the process of finding the stem of word based on the linguistic lexicons. The first step of this stemming is pre-processing. The function of the pre-processing is to identify the sentences boundaries, to split the running text into tokens so that it can be fed into morphological analyser and parser processing. This step is to remove redundant and misspelled space. It also to resolve the orthographic variation in Arabic writing which can be change or unchanged the meaning but always affect the NLP system, such as:

- *Uses of* ي *vs.* ى (*Yeh vs. Alif Maqsura*)
- *Uses of* ه *vs.* ة (*Heh vs. Taa Marbuta*)
- *Initial Alif Variations*: (ا أ إ ٱ آ)

The second step is the named entity recognition. The main purpose of this step is to identify Arabic names using some heuristic also some lists of special verbs that are identified as introducing person names and descriptives that are identified to be linked to person names. The general idea behind this process is that most of the Arabic names are real words that frequently used. The process of identification them early prevents the system from manipulating them as other Arabic words. The third step is multi-word expression (MWE) identification. Multi-word Expressions (MWEs) are two or more words that act like a single word syntactically and semantically [4]. MWE cover expressions that are traditionally classified as idioms (e.g. down the drain), prepositional 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). According to [10], the collocation is defined as "the two or more words which appear together and always seems as comrades". The final step of this stemmer is identifycation of stem. This step is to use the Arabic lexicons (suffixes, prefixes, and stems) for extracting the stem of the word. These procedures require some linguistic information of Arabic such as, Arabic stems, prefixes, suffixes, and clitics. From LDC, the lexicon file stems that were collected by [14] have been selected as the database of the current system. The stems in this file are in Romanic, for this reason, they need to transliterate to Arabic with ignoring diacritics before using them. Before using this file, all stems have been transliterated from Romanic to Arabic and collapsed the entries that have the same stem and same morphological category. This step consists of the following procedures:

1) Select the stems from lexicon that are contained in the word.

2) Match the stems to the word in order to identify the affixes (prefix and suffix) of word.

3) If the identified prefix exists in the Arabic prefixes, and the identified suffix exists in the Arabic suffixes, then check the contradiction of affixes.

4) Select the stem that has the shortest length.

This step is to look for all possible stems for the word that are contained in the word. The Arabic word may have more than two stems that can construct the word. From the stems lexicon, all stems that construct the word and their length more than 2 have been selected as the candidates of stem. For example, the word "التطبيقية" has the following stems: "تطبيق", "تطبيقي", and "طبي". **Table 1** shows all possible segmentations of the word "التطبيقية".

From **Table 1**, the procedure three checks whether the prefix and suffix exist in the prefix and suffix lexicons respectively. The prefix and suffix that do not exist in the lexicon will be ignored. After that, only the stems that their affixes have no contradiction of affixes will be selected as the real stems. For example, the stem "طبي" will be ignored because its prefix and suffix do not exist in the prefix and suffix lexicons respectively. The procedure four is to select one stem from the candidates of stem for the word that have shortest length. For example, in **Table 1**, the candidates of stems for the word "التطبيقية"

Table 1. Segmentations of the word "التطبيقية".

prefix	stem	suffix
ال	تطبيق	ية
ال	تطبيقي	ة
الت	طبي	قية

are "تطبيق", and "تطبيقي". The stem "تطبيق" will be selected because it has shortest length.

4. Results

In our experiment, we have used the Arabic corpus. Our corpus is an electronic corpus of Modern Standard Arabic that was collected from online Arabic newspaper archives. This corpus includes ten documents with different sizes (the number of words). **Table 2** provides the numerical details about the Arabic corpus used in the method for word stemming.

5. Evaluation

The main objective of this experiment is to evaluate the enhanced stemmer in ten documents that compose the corpus. The enhanced stemmer is applied on each document to extract the stem of words and compute the accuracy of this stemmer.

The experiment shows that the ten documents can easily be combined into a single table, which then provides a complete picture of the differences between the accuracy of the stemmer. **Table 3** depicted that the highest accuracy value (97.12%) was achieved by the second document with number of words equal to 7146. In contrast, the lowest accuracy value (95.56%) was achieved by the hybrid steemer in the sixth document with number of words equal to 3649.

Also, the light stemmer and dictionary-based stemmer are applied on the same corpus to extract the stem of words. This experiment is to compare the precision values of the three stemmers (light, dictionary-based, and enhanced stemmer). **Table 4** contains all the documents in the corpus with the accuracy values for each stemmer in each document.

The evaluation graph (**Figure 1**) can present the same information in the evaluation table in more intuitive and readable way. In any evaluation graph, the *x*-axis represents all documents; the *y*-axis gives the corresponding accuracy. The corresponding accuracy for the stemmers can be determined from the intersection of each vertical line with the respective precision graphs, allowing the reader to reconstruct the detailed information provided in the evaluation table.

Table 2. Statistics on the corpus used in stemming.

Statistics	Value
Size (KB)	242
Documents	10
Words	71.935
Sentences	3.596

Table 3. The accuracy of enhanced stemmer.

Doc	Words	Stop-words	Names	MWE	Not Stemmed	Accuracy
1	1317	298	62	17	43	0.9674
2	7146	1266	210	115	206	0.9712
3	9183	2310	215	70	320	0.9652
4	7079	1376	227	93	265	0.9626
5	3879	733	139	31	160	0.9588
6	3649	692	118	33	162	0.9556
7	7172	1520	249	87	257	0.9642
8	2034	439	47	20	75	0.9631
9	12657	2541	440	100	484	0.9618
10	17114	3361	561	149	689	0.9597

Table 4. The evaluation graph for three stemmer on ten documents.

No of Dec.	LS	DBS	ES
1	0.86	0.8747	0.9674
2	0.8756	0.9022	0.9712
3	0.8551	0.9014	0.9652
4	0.8706	0.8874	0.9626
5	0.8361	0.8711	0.9588
6	0.8415	0.8797	0.9556
7	0.8567	0.8989	0.9642
8	0.8531	0.8835	0.9631
9	0.8599	0.8823	0.9618
10	0.8502	0.8821	0.9597

6. Discussion

Figure 1 shows the enhanced stemmer clearly outperforms the others stemmers (light stemmer and dictionary-based stemmer). It achieved the highest accuracy values in all documents in the corpus. The accuracy values of enhanced stemmer had been increased in all documents in the corpus when they compared with the accuracy values in light and dictionary-based stemmer. This improvement of accuracy values is due to the solving the problems in light stemmer and dictionary-based stemmer. The irregular words that are exempted from stemming in the light stemmer had been stemmed by the dictionary-based stemmer; in contrast, the words that are not found in the lexicon of the dictionary-based stemmer have been stemmed by light stemmer. Furthermore, the dictionary-based stemmer is better than the light stemmer that achieved the highest accuracy values in all documents

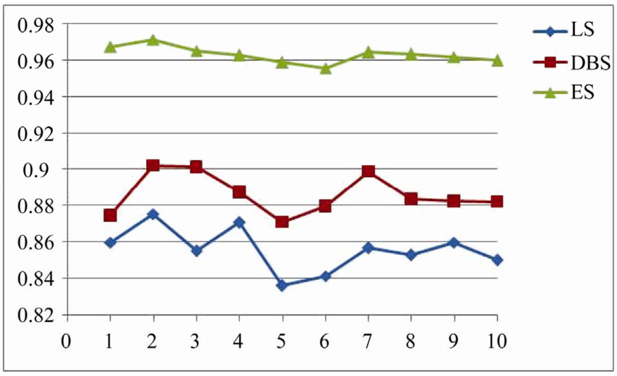

Figure 1. The evaluation graph for three stemmer on ten documents.

in the corpus, with accuracy ranging from 87.11% in the fifth document to 90.22% in the second document. In the evaluation of stemmers, the accuracy value of the stemmer is affected by the following factors:

1) The type of approach: the stemmers have different precision values with different types of approaches in the same data.

2) The corpus: the size and composition of the corpus that is used for evaluation plays an important role in increasing or decreasing the precision values for the stemmers.

3) The pre-processing: this includes some linguistic tools such as the tokenization, identification of Arabic stop-words, named entity recognition, and handling of Arabic multi-word expressions. These linguistic tools are used to reduce the ambiguity of words in order to increase the accuracy and effectiveness of the stemmer.

7. Conclusions

In this study, we have presented the enhanced stemming for extracting the stem and root of Arabic words. The enhanced stemming was designed to overcome the disadvantages of the light stemming and dictionary-based stemming. The problem of the broken (irregular) plurals for nouns and irregular verbs that cannot be solved by the light stemmer has been identified by the dictionary-based stemmer. In contrast, the words that cannot be stemmed in the dictionary-based stemmer because they are not found in the lexicon of Arabic stems have been handled by the light stemmer. In order to evaluate the enhanced stemmer, we applied our method for an in-house collected corpus from Arabic newspaper archives. In our experiment, the average of accuracy in enhanced stemmer on the corpus is 96.29%. The accuracy values of enhanced stemmer had been increased in all documents in the corpus when they compared with the accuracy values in light stemmer (85.5%) and dictionary-based stemmer (88.63%). The accuracy value of stemmer depends on many factors, including the type of stemming approach, the size and composition of the corpus, and pre-processing (such as tokenization, identification of Arabic stop-words, named entity recognition, and handling of Arabic multi-word expressions). The enhanced stemming method that had been demonstrated for extracting the root and stem of Arabic words can be straightforwardly expanded to identify the linguistic category of the word.

REFERENCES

[1] Al. Hajjar, M. Hajjar and K. Zreik, "A New System for Evaluation of Arabic Root Extraction Methods," *Proceedings of the 5th International Conference on Internet and Web Applications and Services*, *ICIW*, Barcelona, Spain, 9-15 May 2010, pp. 506-512.

[2] E. Al-Shammari and J. Lin, "A Novel Arabic Lemmatization Algorithm," *Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data*, Singapore, 24 July 2008.

[3] B. Al-Salemi and M. J. Ab Aziz, "Statistical Bayesian Learning for Automatic Arabic Text Categorization", *Journal of Computer Science*, Vol. 7, No. 1, 2011, pp. 39-45. doi:10.3844/jcssp.2011.39.45

[4] K. R. Beesley and L. Karttunen, "Finite-State Morphology: Xerox Tools and Techniques," CSLI, Stanford, 2003.

[5] K. Shaalan, M. Magdy and A. Fahmy, "Morphological Analysis of Ill-Formed Arabic Verbs in Intelligent Language Tutoring Framework," *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference*, 19-21 May 2010, pp. 277-282.

[6] M. A. Attia, "An Ambiguity Controlled Morphological Analyzer for Modern Standard Arabic Modeling Finite State Networks," *Proceedings of the Challenge of Arabic for NLP/MT Conference*, The British Computer Society, London, 2006.

[7] A. Boudlal, R. Belahbib, A. Lakhouaja and A. Mazroui, "A Markovian Approach for Arabic Root Extraction," *The International Arab Journal of Information Technology*, Vol. 8, No. 1, 2009, pp. 13-20.

[8] M. Sawalha and E. Atwell, "Adapting Language Grammar Rules for Building Morphological Analyzer for Arabic Language," *Proceedings of the Workshop of Morphological Analyzer Experts for Arabic language*, organized by Arab League Educational, 2009.

[9] R. Sonbol, N. Ghneim and M. S. Desouki, "Arabic Morphological Analysis: A New Approach. Information and Communication Technologies: From Theory to Applications," *The 3rd International Conference on Information & Communication Technologies*: *From Theory to Applications*, 7-11 April 2008, pp. 1-6.

[10] A. A. Mohd Juzaiddin, A. Fatimah, A. A. Abdul Azim and M. Ramlan, "Pola Grammar Technique to Identify Subject and Predicate in Malaysian Language," *The Second International Joint Conference on Natural Language Processing*, 11-13 October 2005, pp. 185-190.

[11] A. M. Saif and M. J. A. Aziz, "An Automatic Collocation Extraction from Arabic Corpus," *Journal of Computer Science*, Vol. 7, No. 1, 2011, pp. 6-11.

[12] K. Taghva, R. Elkoury and J. Coombs, "Arabic Stemming without a Root Dictionary," *International Conference on Information Technology*: *Coding and Computing* (*ITCC*' 05), 4-6 April 2005, pp. 152-157.

[13] R. Alshalabi, "Pattern-Based Stemmer for Finding Arabic Roots" *Asian Network for Scientific Information Technology Journal*, Vol. 4, No. 1, 2005. pp. 38-43.

[14] T. Buckwalter, "Issues in Arabic Orthography and Morphology Analysis," *The Workshop on Computational Approaches to Arabic Script-Based Languages*, COLING Geneva, 2004, pp. 31-34.