Published at the 6th *International Conference on Electrical and Computer Systems (EECS'10)*, Nov 25-26, 2010, Lefke, North Cyprus.

Arabic Morphological Tools for Text Mining

Motaz K. Saad
Faculty of Information Technology
Islamic University of Gaza
Gaza, Palestine
e-mail msaad@iugaza.edu.ps

Wesam Ashour
Computer Engineering Department
Islamic University of Gaza
Gaza, Palestine
e-mail washour@iugaza.edu.ps

***Abstract*—Arabic Language has complex morphology; this led to unavailability to standard Arabic morphological analysis tools until now. In this paper, we present and evaluate existing common Arabic stemming / light stemming algorithms, we also implement and integrate Arabic morphological analysis tools into the leading open source machine learning and data mining tools, Weka and RapidMiner.**

Keywords: Arabic language, Arabic morphological tools, Arabic stemming / light stemming.

1. INTRODUCTION

Arabic Language is the 5th widely used languages in the world. It is spoken by more than 422 million people as a first language and by 250 million as a second language [4]. Arabic has 3 forms; Classical Arabic (*CA*), Modern Standard Arabic (*MSA*), and Dialectal Arabic (*DA*). *CA* includes classical historical liturgical text, MSA includes news media and formal speech, and *DA* includes predominantly spoken vernaculars and has no written standards. Arabic alphabet consists of the following 28 letters (أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي) in addition, the Hamza (ء). There is no upper or lower case for Arabic letters like English letters. The letters (أ و ي) are vowels, and the rest are constants. Unlike Latin-based alphabets, the orientation of writing in Arabic is from right to left.

Table 1: Diacritics

Double Constant	No Vowel	Nunation			Vowel		
بّ /bb/	بْ /b/	بٍ /bin/	بٌ /bun/	بً /ban/	بِ /bi/	بُ /bu/	بَ /ba/

حقوق الانسان
حقـــــوق الانســـــان
حقـــــــــــوق الانســـــــــــان
حقـــــــــــــــــوق الانســـــــــــــــــان

Fig. 1: Tatweel (kasheeda)

The Arabic script has numerous diacritics, including I'jam (إعجام), consonant pointing, and tashkil (تشكيل), supplementary diacritics. The latter include the ḥarakat (حركات, singular haraka حركة), vowel marks. The literal meaning of tashkil is "forming". As the normal Arabic text does not provide enough information about the correct pronunciation, the main purpose of tashkil (and ḥarakat) is to provide a phonetic guide or a phonetic aid; i.e. show the correct pronunciation (double the word in pronunciation or to act as short vowels). The ḥarakat, which literally means "motions", are the short vowel marks [3]. Arabic diacritics include Fatha, Kasra, Damma, Sukūn, Shadda, and Tanwin. The pronunciations of diacritics aforementioned are presented in Table 1. Arabic words may also have Tatweel or kasheeda as shown in figure 1.

Arabic words have two genders, masculine (مذكر) and feminine (مؤنث); three numbers, singular (مفرد), dual (مثنى), and plural (جمع); and three grammatical cases, nominative (الرفع), accusative (النصب), and genitive (الجر). A noun has the nominative case when it is subject (فاعل); accusative when it is the object of a verb (مفعول); and the genitive when it is the object of a preposition (مجرور بحرف جر). Words are classified into three main parts of speech, nouns (اسماء) (including adjectives (صفات) and adverbs (ظروف)), verbs (افعال), and particles (ادوات).

The rest of this paper is organized as follows: section 2 presents the complexity of Arabic language, section 3 evaluates common Arabic stemmer / light stemmer, and presents our implementation and integration of Arabic morphological tools into leading open source machine learning and data mining tools. Finally section 4 draws the conclusion.

2. COMPLEXITY OF ARABIC LANGUAGE

Arabic is a challenging language for a number of reasons:

- Orthographic (الاملاء) with diacritics is less ambiguous and more phonetic in Arabic, certain combinations of characters can be written in different ways [3].
- Arabic language has short vowels which give different pronunciation. Grammatically they are required but omitted in written Arabic texts [4].
- Arabic has a very complex morphology as compare to English language [1, 2, 6, 7, 8, 11, 12, 15, 18, 18, 19].
- Synonyms are widespread. Arabic is a highly inflectional and derivational language [6, 7, 18].

- Lack of publically freely accessible Arabic Corpora [18, 19].
- Lack of Arabic digital contents [18, 19].

In the following, we shall discuss these points in details.

Word meanings: It is possible to identify the different meanings associated with a word, due to one word may have more than one meaning in different contexts.. Table 2 shows the Arabic word (قلب) which has 3 meaning as a noun.

Table 2: The meaning of word (قلب) as a noun

Word meaning	Sentence
core	في **قلب** الاحداث
heart	اجرى عملية **قلب** مفتوح
center, middle	الكرة في **قلب** الملعب

Variations in lexical category: One word may have more than lexical category (noun, verb, adjective, etc.) in different contexts as shown in Table 3. Morphological analysis of a given corpus includes investigating word frequency of a word as a lexical category.

Table 3: The Lexical Category of word (عين)

Word meaning	Word Category	Sentence
Ain	Proper-Noun	**عين** جالوت
wellspring	Noun	**عين** الماء
eye	Noun	**عين** الانسان
delimitate/be delimitate	Verb/passive Verb	**عين** وزيرا للخارجية

Synonyms: Languages have many words that are considered synonymous. Through a given corpus, the researchers can use morphological analysis tools to know synonyms of a word, the frequency of each word of those synonyms and which one of them is more common. Examples of synonyms in Arabic are (بذل منح اعطى وهب) which means (give), (اسرة عائلة) which means (family), and (فصل صف) which means (classroom).

The word form according to its case: The form of some Arabic words may change according to their case modes (nominative, accusative or genitive). For instance, the plural of word (مسافر) which means (traveler) may be in the form (مسافرون) in the case of nominative (مرفوعة) and the form (مسافرين) in the case of accusative/genitive (منصوبة/مجرورة). Arabic light stemming can handle these cases.

Morphological characteristics: An Arabic word may be composed of a stem plus affixes and clitics. The stem consists of a consonantal root (جذر صحيح) and a pattern morpheme (اصغر كلمة ذات معنى). The affixes include inflectional markers (علامات او حركات اعرابية) for tense, gender, and/or numbers. The clitics include some prepositions (حروف جر), conjunctions (حروف العطف), determiners (محددات), possessive pronouns (ضمائر الملكية) and pronouns (ضمائر). The clitics attached to the beginning of a stem are called proclitic and the ones attached to the end of it are called enclitics. Most Arabic morphemes are defined by three consonants, to which various affixes can be attached to create a word. For example, from the tri-consonant "ktb" (كتب), we can inflect (يصرف) several different words concerning the idea of writing as (wrote كَتب), (book كِتاب), (the book الكِتاب), (books كُتُب), (he writes يكْتُب), (writer كاتِب), (library مكْتَبَة). Moreover an Arabic word may correspond to several English words. Another example is the Arabic word (وبنفوذها) and its equivalence in English "and with her influences". This makes segmentation of Arabic textual data different and more difficult than Latin languages.

Affixes set in Arabic are shown in Table 4, and Arabic patterns (الأوزان) and roots are shown in Table 5. The word (علم) may give various meanings by adding different affixes (prefixes, infixes, or suffixes) as shown in Table 6. Other morphological variations example is the word (يذهب) which means (go) are presented in Table 7.

Table 4: Affix set in Arabic Language

Affixes in Arabic	Examples
Prefixes of length 3	ولل ، وال ، كال ، بال
Length 2 prefixes	ال ، لل
Length 1 prefixes	ل ، ب ، ف ، س ، و ، ى ، ت ، ن ، ا
Length 3 suffixes	تمل ، همل ، تان ، تين ، كمل
Length 2 suffixes	ون ، ات ، ان ، ين ، تن ، كم ، هن ، نا ، يا ، ها ، تم ، كن ، ني ، وا ، ما ، هم
Length 1 suffixes	ة ، ه ، ي ، ك ، ت ، ا ، ن

Table 5: Arabic Patterns and Roots

Arabic Pattern and roots (الأوزان)	Examples
Length 4 pattern	فاعل فعلة فعال مفعل
Length 5 pattern and length 3 roots	تفاعل افتعل افعال فعالة فعلان فعولة تفعلة تفعيل مفعلة مفعول فاعول فواعل مفاعل مفعيل افعلة فعائل منفعل مفتعل فاعلة مفاعل فملاع يفتعل تفتعل فعلالي انفعل
Length 5 pattern and length 4 roots	تفعلل افعلل مفعلل فعللة فعلان فعالل
Length 6 pattern and length 3 roots	استفعل مفاعلة افتعال افعوعل انفعال مستفعل
Length 6 pattern and length 4 roots	افتعلل افعلال متفعلل

Table 6: Versions of the word (علم) and its meaning when adding affixes

Meaning	Suffix	Infix	Prefix	Word
Scientific	ية	***	***	علمية
Learned us	تنا	***	***	علمتنا
His science	ه	***	***	علمه
Scientists	اء	***	***	علماء
Teaching	***	ي	ت	تعليم
Sciences	***	و	***	علوم
Informative	يه	ا	است	استعلامية

Stemming usually used to convert words to root form, it dramatically reduces the complexity of Arabic language morphology by reducing the number of feature / keywords in corpora. The reason for using stemming as feature / keywords reduction technique is that all morphology of words mostly has the same context meaning, but the case is not always true. Table 8 shows some of these cases. There is another approach for morphology reduction that just removes

affixes and does not convert the word to bas/root form. This approach is called light stemming [12, 18]. More details in section 3.

Table 7: Morphological variation of word (ذهب)

verb	time	# of participants	Gender of subjects
ذهب	Past	1	Male
ذهبت	Past	1	Female
ذهبا	Past	2	Male
ذهبتا	Past	3	Female
ذهبوا	Past	3 or more	Male
ذهبن	Past	3 or more	Female
يذهب	Present	1	Male
تذهب	Present	1	Female
سيذهب	Future	1	Male
ستذهب	Future	1	Female
سيذهبوا	Future	3 or more	Male
سيذهبن	Future	3 or more	Female

Table 8: Different meaning of morphology of the same root in Arabic

Meaning	Root	Word
Class room **Apartheid**	فصل	**الفصل** الدراسي **الفصل** العنصري
Goes out of house **Graduate from university**	خرج	**يخْرج** من البيت **تخْرج** من الجامعة
The fisherman twist the cord **The student argued with the teacher**	جدل	**جدل** الصياد الحبل **جادل** الطالب المدرس
He focuses the arrow **The man lost his mind**	صوب	انه **يصوب** السهم فقد الرجل **صوابه**

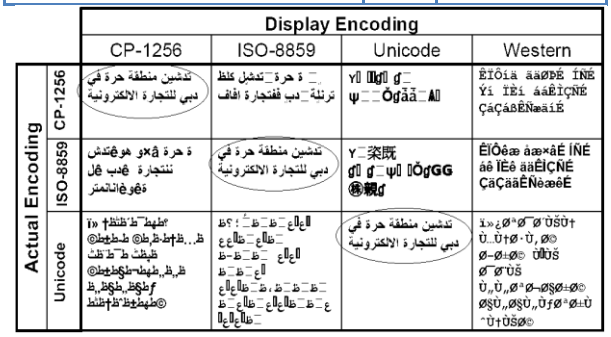

Fig. 2: Arabic Encoding Problem

Table 9: Unicode vs. cp-1256 Arabic windows encoding

Unicode	CP-1256 Arabic windows
Becoming the standard more and more	Commonly used
2-byte characters	1-byte characters
Widely supported input/display	Widely supported input/display
Supports extended Arabic characters	Minimal support for extended Arabic characters
Multi-script representation	bi-script support (Roman/Arabic)
Supports presentation forms (shapes and ligatures)	Tri-lingual support: Arabic, French, English (ala ANSI)

Encoding Problem: Arabic Language has display Problems (encoding issues) because it has different encoding according to machine platform. Figure 2 shows the problem of using incorrect encoding where all circled cells are displayed correctly while the other cells are displayed incorrectly. Text preprocessing, mining, and information retrieval with incorrect encoding may lead to incorrect results. Table 9 presents the characteristics of two common Arabic encoding systems; *Unicode* and code page 1256 *CP-1256* Arabic windows.

3. Arabic Morpholigical Tools

In linguistics, morphology is the identification, analysis and description of the structure of morphemes and other units of meaning in a language like words, affixes, and parts of speech and intonation/stress, implied context (words in a lexicon are the subject matter of lexicology) [10, 11]. Morphological typology represents a way of classifying languages according to the ways by which morphemes are used in a language from the analytic that use only isolated morphemes, through the agglutinative ("stuck-together") and fusional languages that use bound morphemes (affixes), up to the polysynthetic, which compress lots of separate morphemes into single words [10, 11].

While words are generally accepted as being (with clitics) the smallest units of syntax, it is clear that in most (if not all) languages, words can be related to other words by rules (grammars). For example, English speakers recognize that the words dog and dogs are closely related — differentiated only by the plurality morpheme "-s," which is only found bound to nouns, and is never separate. Speakers of English (a fusional language) recognize these relations from their tacit knowledge of the rules of word formation in English. They infer intuitively that dog is to dogs as cat is to cats; similarly, dog is to dog catcher as dish is to dishwasher (in one sense). The rules understood by the speaker reflect specific patterns (or regularities) in the way words are formed from smaller units and how those smaller units interact in speech. In this way, morphology is the branch of linguistics that studies patterns of word formation within and across languages, and attempts to formulate rules that model the knowledge of the speakers of those languages [10, 11].

Terms have many morphological variants that will not be recognized by term matching algorithm without additional text processing. Stemming algorithms are needed in many applications such as natural language processing, compression of data, and information retrieval systems. In most cases, these variants have similar semantic interpretation and can be treated as equivalence. Thus, stemming algorithm can be employed to perform term reduction to a root form [10, 11].

In general, most of Arabic morphological tools face a problem with diacritics because most of them remove (normalize) diacritics. For example, the Arabic word (ذَهَبَ) which means (went) has identical form (without diacritics) to word (ذَهَبٌ) which means (gold). Diacritics distinguish between them, but unfortunately, most of Arabic morphological tools remove them as a first step [10, 11, 12].

For Arabic Language, there are two different morphological analysis techniques; stemming and light stemming. Stemming reduces words to their stems [6, 7, 19]. Light stemming, in contrast, removes

common affixes from words without reducing them to their stems [11]. Stemming would reduce the Arabic words (الكتاب الكاتب المكتبة) which mean (the library), (the writer), and (the book) respectively, to one stem (كتب), which means (write).

The main idea for using light stemming [6, 7, 12] is that many word variants do not have similar meanings or semantics although these word variants are generated from the same root. Thus, root extraction algorithms affect the meanings of words. Light stemming aims to enhance feature/keyword reduction while retaining the words 'meanings. It removes some defined prefixes and suffixes from the word instead of extracting the original root [6, 7]. Formally speaking, the aforementioned Arabic words (الكتاب الكاتب المكتبة) which mean (the library), (the writer), and (the book) respectively, belong to one stem (كتب) despite they have different meanings. Thus, the stemming approach reduces their semantics. The light stemming approach, on the other hand, maps the word (الكتاب) which means (the book) to (كتاب) which means (book), and stems the word (الكاتب) which means (the writer) to (كاتب) which means (writer). Another example for light stemming is the words (المسافرون المسافرين) which mapped to word (مسافر). Light stemming keeps the words' meanings unaffected. We previously described in section 2 that there are many words morphology have different meaning despite they have the same root. Figure 3 shows the steps of Arabic light stemming [12]. Arabic light stemmer is implemented in Apache Lucene as a standard Arabic light stemmer.

Stemming algorithm by Khoja [11] is one of well know Arabic Stemmers. Khoja's stemmer removes the longest suffix and the longest prefix. It then matches the remaining word with verbal and noun patterns, to extract the root. The stemmer makes use of several linguistic data files such as a list of all diacritic characters, punctuation characters, definite articles, and stopwords.

1. Normalize word
   - Remove diacritics
   - Replace إ أ آ with ا
   - Replace ة with ه
   - Replace ى with ي
   - Remove diacritics
2. Stem prefixes
   - Remove Prefixes: الـ ، والـ ، بالـ ، كالـ ، فالـ ، للـ ، و
3. Stem suffixes
   - Remove Suffixes: ها ، ان ، ات ، ون ، ين ، ية ، ه ، ي

Fig. 3: Arabic Light Stemming Algorithm Steps

However, the Khoja stemmer has several weaknesses [16]. First, the root dictionary requires maintenance to guarantee newly discovered words are correctly stemmed. Second, the Khoja stemmer replaces a weak letter with (و) which occasionally produces a root that is not related to the original word. For example, the word (منظمات) which mean (organizations) is stemmed to (ظمآ) which means (he was thirsty) instead of (نظم). Here the Khoja stemmer removed a part of the root when it removed the prefix and then added a hamza at the end. Third, by following a certain order of affixes, the Khoja stemmer will in some cases fail to remove all of them. For example, the terms (تستغرق) and (ركبتيه) are not stemmed although they are respectively derived from the two regular roots (غرق) and (ركب). Algorithm steps of Khoja Arabic stemmer [11] is described in Figure 3.

1. Remove diacritics
2. Remove stopwords, punctuation, and numbers.
3. Remove definite article (ال)
4. Remove inseparable conjunction (و)
5. Remove suffixes
6. Remove prefixes
7. Match result against a list of patterns.
   - If a match is found, extract the characters in the pattern representing the root.
8. Match the extracted root against a list known "valid" roots
9. Replace weak latters واي with و
10. Replace all occurrences of Hamza ئ ء ؤ with ا
11. Two letter roots are checked to see if they should contain a double character. If so, the character is added to the root.

Fig. 4: Arabic Stemming Algorithm Steps

Al-Shalabi, et. al. [2] developed a root extraction algorithm (tri-literal root extraction) which does not use any dictionary. It depends on assigning weights for a word's letters multiplied by the letter's position, Consonants were assigned a weight of zero and different weights were assigned to the letters grouped in the word (سألتمونيها) where all affixes are formed by combinations of these letters. The algorithm selects the letters with the lowest weights as root letters.

Sawalhi and Atwell [15] evaluated Arabic Language Morphological Analyzers and Stemmers. They reported Khoja stemmer achieved the highest accuracy then the tri-literal root extraction algorithm. The majorities of words have a tri-lateral root, in fact between 80 and 85% of words in Arabic are derived from tri-lateral roots [1, 8]. The rest have a quad-letter root, penta-letter root or hexa-letter root. Khoja stemmer works accurately for tri-literal roots, this why it achieved the highest accuracy. Sawalhi and Atwell also reported that most stemming algorithms are designed for information retrieval systems where accuracy of the stemmers is not important issue. On the other hand, accuracy is vital for natural language processing. The accuracy rates show that the best algorithm failed to achieve accuracy rate of more than 75%. This proves that more research is required. We cannot rely on such stemming algorithms for doing further research as Part-of-Speech tagging and then Parsing because errors from the stemming algorithms will propagate to such systems [15].

3.1 Text Preprocessing tools

We implement and integrate Arabic stemming and light stemming algorithms, described in Figures 3, and 4 respectively, to the leading open source Machine Learning and Data Mining tools, WEKA and RapidMiner. We adopt Arabic stopwords list from [5] for stopwords removal. The complete package of integration is available publically at [14].

WEKA (Waikato Environment for Knowledge Analysis) [9] is a popular suite of machine learning software written in Java, developed at the University of Waikato. It is free software available under the GNU General Public License. WEKA provides a large

collection of machine learning algorithms for data pre-processing, classification, clustering, association rules, and visualization, which can be invoked through a common Graphical User Interface. A screenshot of Arabic stemmer / light stemmer integrated to Weka is depicted in Figure 5.

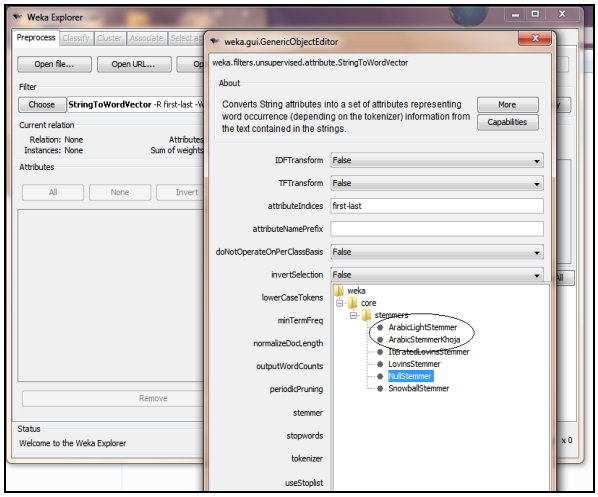

Figure 5: Weka Arabic Stemmer/Light Stemmer

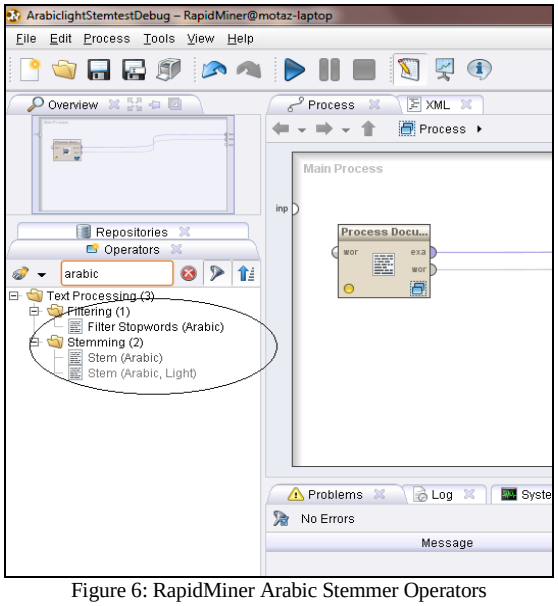

Figure 6: RapidMiner Arabic Stemmer Operators

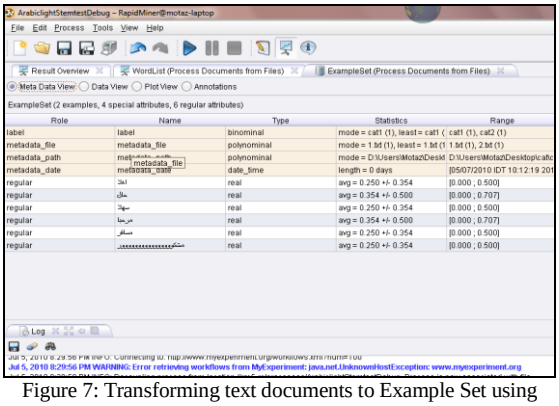

Figure 7: Transforming text documents to Example Set using RapidMiner

RapidMiner (formerly YALE (Yet Another Learning Environment)) is an environment for machine learning and data mining experiments. It allows experiments to be made up of a large number of arbitrarily nestable operators. Operators are described in XML files which are created with RapidMiner' s graphical user interface. RapidMiner is used for both research and real-world data mining tasks [13]. RapidMiner provides more than 1,000 operators for all main machine learning procedures, including input and output, and data preprocessing and visualization. It is written in the Java programming language and therefore can work on all popular operating systems. It also integrates learning schemes and attributes evaluators of the Weka learning environment [13]. We implemented and contributed 3 operators to RapidMiner text plugin; Arabic Stemmer, Arabic Light Stemmer, and Arabic stopwords removal operator. The contribution is available publically within text processing RapidMiner plugin. Figure 6 shows a screenshot of the three operators. Figure 7 shows the process of transforming text documents to record using RapidMiner. Figure 8 shows the resulting wordlist (dictionary).

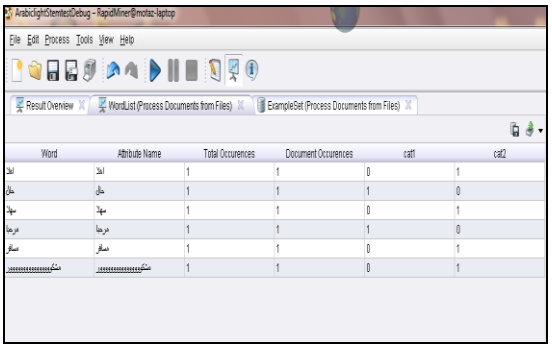

Figure 8: Transforming text documents to word list using RapidMiner

Arabic stemming/light stemming implementation and integration into WEKA and RapidMiner were used by Saad [18] to address the impact of text preprocessing on Arabic text classification.

4. CONCLUSION

The reason for unavailability of standard Arabic morphological analysis tools is the complex nature of Arabic language. Thus, more researches in the field are needed. There is a lack of Arabic morphological analysis tools. In this paper, we evaluated common existing Arabic stemmer/light stemmer. We also implement and integrate Arabic morphological analysis tools into leading open source data mining / machine learning tools.

REFERENCES

[1]. Al-Fedaghi, S., Al-Anzi, F.: *A new algorithm to generate Arabic root-pattern forms.* In Proc. of the 11th National Computer Conf. King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia, (pp. 04–07). 1989.

[2]. Al-Shalabi, R., Kanaan, G., Al-Serhan, H.: *New approach for extracting Arabic roots.* Int. Arab Conf. on Information Technology (ACIT'2003), Egypt. 2003.

[3]. Arabic diacritics - Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Arabic_diacritics

[4]. Arabic language - Wikipedia, the free encyclopedia, http://ar.wikipedia.org/wiki/لغة_عربية
[5]. Arabic Stop words, http://sourceforge.net/projects/arabicstopwords, 2010.
[6]. Duwairi R., Al-Refai M., Khasawneh N.: *Feature reduction techniques for Arabic text categorization.* Journal of the American Society for Information Science. Volume 60 Issue 11, pp. 2347 - 2352. 2009.
[7]. Duwairi R., Al-Refai M., Khasawneh N.: *Stemming Versus Light Stemming as Feature Selection Techniques for Arabic Text Categorization*. 4th Int. Conf. on Innovations in Information Technology. IIT '07. pp.: 446 - 450. 2007.
[8]. Eldin, S.: *Development of a computer-based Arabic Lexicon.* In The Int. Symposium on Computers & Arabic Language (ISCAL) Riyadh, KSA. 2007.
[9]. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.: *The WEKA Data Mining Software: An Update.* SIGKDD Explorations, Vol. 11, No. 1, 2009.
[10]. Hill T., Lewicki P.: *STATISTICS Methods and Applications*. StatSoft, Tulsa, OK. 2007.
[11]. Khoja S., Garside R.: *Stemming Arabic text.* Computer Science Department, Lancaster University, Lancaster, UK, 1999.
[12]. Larkey L, Ballesteros L, Connell M.: Light Stemming for Arabic Information Retrieval. Arabic Computational Morphology, book chapter, Springer. 2007.
[13]. Mierswa I., Wurst M., Klinkenberg R., Scholz M., Euler T.: *YALE: Rapid Prototyping for Complex Data Mining Tasks*. in Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD-06), 2006.
[14]. Motaz K. Saad: *Open Source Arabic Language and Text Mining Tools.* 2010. http://sourceforge.net/projects/ar-text-mining
[15]. Sawalha, M, Atwell, E.: *Comparative evaluation of Arabic language morphological analyzers and stemmers.* In Proc. of COLING 2008 22nd Int. Conf. on Computational Linguistics. 2008.
[16]. Taghva, K., Elkhoury, R., Coombs, J.: *Arabic stemming without a root dictionary*. Information Technology: Coding and Computing, ITCC, Vol. 1, pp 152 – 157, 2005.
[17]. Saad M. K., Ashour W., *Arabic Text Classification Using Decision Trees*, Proceedings of the 12th international workshop on computer science and information technologies CSIT'2010, Moscow – Saint-Petersburg, Russia, 2010.
[18]. Saad M. K., *The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification*, MSc. Thesis Dissertation, Computer Engineering Dept., Islamic University of Gaza, Palestine, 2010.
[19]. Saad M. K., Ashour W., *OSAC: Open Source Arabic Corpora*, 6th ArchEng Int. Symposiums, EEECS'10 the 6th Int. Symposium on Electrical and Electronics Engineering and Computer Science, European University of Lefke, Cyprus, 2010.