

Using Arabic skelton morphology and Maximum Entropy for Arabic Document classification

Abstract- The morphology of Arabic plays an important role of computational natural language processing systems. The rich morphology, and the complexity of word formation all contribute to making morphological approaches to Arabic very challenging .In this paper, we present a new method for Arabic document classification using maximam entropy and morphological derivation for Arabic words. In this paper, maximum entropy and Arabic word derivative morphology for text classification by estimating the conditional distribution of the class variable given the document. Using these derivatives we can find a related words in the document which contains words and its derivatives. The proposed approach is designed for vowel and unvowel Arabic document.

Keywords- Document classification; Arabic information retrieval; Arabic morphology; maximum entropy.

I. INTRODUCTION

Document Classification (DC) is the process of groping similer document in the same class or field. DC has rapidly grown into a major research area [1,2 ,3,9, 10, 15, 18, 20, 27, 28] as illustrated by the Document Understanding Conference (DUC) and Text Analysis Conference (TAC) series. Arabic classification systems still not as sophisticated and as reliable as those developed for other languages like English. This is due to the rich and complex of Arabic morphology. The earlier approach for automatic classification is word extraction, where key sentences from the input documents are selected based on some of the features. Some information in the text is useful for sentence extraction such as, word frequency; title of the text; position of the sentence; a cue phrase in the text and cohesion in text. Stands to reason the important sentences not only contain frequently words but also that's strongly related to the documentary field. The degree of the importance of sentences depended on how its words related to a document.

The Arabic morphology considers special challenge to computational natural language processing system. The rich morphology of Arabic and the more complex word formation all contribute to produces Arabic IR researches depends on Arabic morphology [6, 8, 11, 12, 16, 19, 22, 25]. Arabic morphology becomes an integral part of many Arabic information retrieval system. Arabic offers special challenges for data driven. An Arabic word consists of a stem with a consonantal root and pattern. Furthermore, it contains affixes and vowels; also sometimes the same root with different vowels stands different meaning. For example, there is a clear sense in which the forms “قَلْبُ: qalbu, which means heart in English”, “قَلَبَ: qalaba, which means turn in English” and “قَلَّبَ: qallaba, which means stir in English” are morphologically related to one another although they do not share in meaning. Most pervious work in AIR

depends on stem [5, 26] stemming is a tool used in IR to combat the vocabulary mismatch problem. These require deleting the vowels and it is a big mistake because many words become the same although they differ in meaning.

In this paper we develop a new technique for Arabic document classification using maximum entropy and Arabic morphological derivatives. To find the derivatives for words, we make to find all real derivatives included in Arabic derivation such that; active, passive and imperative for a verb. In addition, Active Participle, Passive Participle, The Elative, Noun of the Instrument, The Adverb and Intensive Adjective for noun. Then construct a morphological ontology for each word with its all possible derivatives and its relation. For example the word “كَتَبَ, kataba which means in English wrote” has a relation with the word “كَاتِبَ, kaatab which means in English writer” as Active Participle

The reset of this paper is organized as follows; Section 2 defines maximum entropy and how use it document classification. In Section 3 Arabic morphology was discussed. Section 4 our new algorithm to extract all derivatives of FA words will illustrate.

Section5

. Section 6 explains Conclusion and future work.

II.MAXIMUM ENTROPY FOR TEXT CLASSIFICATION

The Principle of Maximum Entropy (ME) is based on the idea that when estimating the probability distribution, you should select that distribution which leaves you the largest remaining uncertainty.

The entropy could be maximized logically. Using the constraint and the fact that the probabilities add up to 1, we expressed two of the unknown probabilities in terms of the third.

Next, the possible range of values of the probabilities was determined using the fact each of the three lies between 0 and 1. Then, these expressions were substituted into the formula for entropy S so that it was expressed in terms of a single probability. Then any of several techniques could be used to find the value of that probability for which S is the largest.

In order to apply maximum entropy to a domain, we need to select a set of features to use for setting the constraints.

For text classification with maximum entropy, we use word counts as our features. More specifically, in this paper for each word-class combination we instantiate a feature as:

$$f_{w,c'}(d, c) = \begin{cases} 0 & \text{if } c \neq c' \\ \frac{N(d,w)}{N(d)} & \text{Otherwise,} \end{cases}$$

where $N(d;w)$ is the number of times word w occurs in document d , and $N(d)$ is the number of words in d . With this representation, if a word occurs often in one class, we would expect the weight for that word-class pair to be higher than for the word paired with other classes. In most natural language tasks using maximum entropy the features are naturally binary features. In text classification, we expect that features accounting for the number of times a word occurs should improve classification.

III. ARABIC MORPHOLOGY

Morphology is the field of linguistics which studied word structure and formation. It consists of inflectional morphology and derivational morphology [4, 7, 17, 21]. Inflectional morphology is defined as the use of morphological methods to form inflected word forms from a lexeme. Inflection word forms indicate grammatical relations between words. On the other hand, derivational morphology is concerned with the derivation of new words from other words using derivational affixes. A morpheme or a root is the smallest unit of language which has a meaning [7, 17]. The morphological processing is needed in natural language processing to handle inflected word forms. Arabic offers a special challenge for derivational morphology.

An Arabic word may be composed of a stem consisting of a consonantal root and a pattern. Furthermore, it contains affixes and vowels. There are 15 triliteral forms, of which at least 9 are common. Within each conjugation pattern, an entire paradigm is found. Two voices (active and passive) and tense (perfect and passive) and five moods (indicative, subjective, jussive, imperative and energetic).

The Arabic vowels

Arabic has 3 short vowels َKasra (E/I), ُFat-ha (A) and ِDamma (U) These are marks that go on top or underneath a letter. If a letter has a vowel, it means that vowel comes after that letter. You can “double” these vowels; this will add the sound of the letter N at the end. This doubling can only happen at the end of a word like َ2 Kasra (EN), ُ2 Fat-ha (AN), ِ2 Damma (UN). If a letter has no vowel after it, we put a special symbol on top of that letter to indicate it, this symbol is called a Sukoon ْ like (َ Yaa (EE) , ُ Aleph (AA) , ِ Waw (UU), َYaa Leen (EI), ُ Waw Leen (AW)). If there is a letter with a Sukoon, then the same letter in the same word again, the two letters will be written as one and a special symbol will be placed on top of the letter to indicate this, This symbol is called the Shadda ّ The vowel of the second letter is placed on top of underneath the Shadda, not on top of underneath the letter itself. The letters Aleph, Waw and Yaa can act as long vowels. The Aleph stretches the Fat-ha vowel to form a long AA sound. The Waw stretches the Damma vowel to form a long OO sound. The Yaa stretches the Kasra vowel to form a long EE sound. So, Aleph must always have a Fat-ha before it. Similarly if Waw is acting as a long vowel, it will have a Sukoon on it and a Damma before it, And if Yaa is acting as a long vowel, it will have a Sukoon on it and a Kasra before it. Waw and Yaa can also act as semi-vowels. Waw can form the semi-vowel AW / OW, as in “Howl”. Yaa can form the semi-vowel EI, as in Hussein. This

will happen if they have a Sukoon and a Fat-ha before them. Several words in Arabic have the same spelling but with different vowels yields to different meaning. For examples, the word (مَالِك: maaleka, which means owner in English) and the word (مَلَك: moleka which means Sovereignty in English). Also, the word (مَيِّت: myaat which means the person who will die but he still alive in english) and the word (مَيِّت: myt which means the person who actually died in English). Also, the word (كَرِه: karah which means in English the hardship which affect rights outside of itself and with force) but the word (كُرِه: koreh which means in English the hardship which affect the human from itself and wish it like Pregnancy and childbirth for women). By this example, we can know that Arabic language is a different language not complex but amazing. For this, we must treat it as a special case, not like other languages. In this paper, proposed a new technique to find a derivation frame for any word in Arabic with its relation for the word. In addition find the real derivatives for words, not useless derivatives. This will be very effective in document summarization.

IV. DERIVATIVES FRAME AND ITS ALGORITHM

Arabic morphology contain more derivatives and this is present, past and imperative verbs for masculine, feminine, dual and plural. Also, actor, object, source, adjective, exaggeration formula, the name of preference, the place name, the time name and instrument name. For example, the verb “فَتَحَ” which means opening in English. The important derivatives as follows:

- 1- Actor the person who opens, “فَاتِح” means opener.
- 2- Object, Signed by the act like “مَفْتُوح” means the opened.
- 3- The place where opening “مَفْتَح”.
- 4- Machine name or the instrument of opening “مِفْتَاح”, which means key in English.
- 5- Derivatives verbs

In this paper we treat with not only the word , but the word with it is derivatives and find the maximum entropy for words and its derivatives .

Algorithm 2: FAD word algorithm

Input: a set of PFA, SPFA words

Output: a set of FAD sets, $FAD = \{$

$w_{1FAD}, w_{2FAD}, \dots, w_{nFAD}\}$. where, w_{iFAD} is a set of derivatives for the word $w_i, 1 \leq i \leq n$. : $*/w_{iFAD} = \{w_{iFAD1}, w_{iFAD2}, \dots, w_{iFADm}\}$

Method

Let $A = \{w_1, w_2, \dots, w_n\}$ is a set of PFA, SPFA

$\forall w_i \in A, 1 \leq i \leq n$

Find w_{iFAD} , where w_{iFAD} is the set of derivatives for each word $w_i \in A$.

{

If $L_{w_i} = 3$, L_{w_i} is the length of a word

let $w_i = C_1V_1C_2V_2C_3V_3$, where C 's is an actual letter or constant and V 's are the vowels.

If $(V_1 = \text{َ} \wedge V_2 = \text{َ})$ then

{

$PrSM = y \text{َ} C_1 \text{َ} C_2 \text{َ} C_3 \text{َ}, y = \text{َ}$

$PrSF = T \text{َ} C_1 \text{َ} C_2 \text{َ} C_3 \text{َ}, T = \text{َ}$

$PrDM = y \circ C_1 \circ C_2 \circ C_3 a \circ, y = \text{ي}, a = \text{ا}$
 $PrDF = T \circ C_1 \circ C_2 \circ C_3 a \circ, T = \text{ت}, a = \text{ا}$
 $PrPM = y \circ C_1 \circ C_2 \circ C_3 wo \circ, y = \text{ي}, wo = \text{و}$
 $PrPF = T \circ C_1 \circ C_2 \circ C_3 N \circ, T = \text{ت}, N = \text{ن}$
 $\}$
Or
 $\{$
 $PrSM = y \circ C_1 \circ C_2 \circ C_3 \circ, y = \text{ي}$
 $PrSF = T \circ C_1 \circ C_2 \circ C_3 \circ, T = \text{ت}$
 $PrDM = y \circ C_1 \circ C_2 \circ C_3 a \circ, y = \text{ي}, a = \text{ا}$
 $PrDF = T \circ C_1 \circ C_2 \circ C_3 a \circ, T = \text{ت}, a = \text{ا}$
 $PrPM = y \circ C_1 \circ C_2 \circ C_3 wo \circ, y = \text{ي}, wo = \text{و}$
 $PrPF = T \circ C_1 \circ C_2 \circ C_3 N \circ, T = \text{ت}, N = \text{ن}$
 $\}$
Or
 $\{$
 $PrSM = y \circ C_1 \circ C_2 \circ C_3 \circ, y = \text{ي}$
 $PrSF = T \circ C_1 \circ C_2 \circ C_3 \circ, T = \text{ت}$
 $PrDM = y \circ C_1 \circ C_2 \circ C_3 a \circ, y = \text{ي}, a = \text{ا}$
 $PrDF = T \circ C_1 \circ C_2 \circ C_3 a \circ, T = \text{ت}, a = \text{ا}$
 $PrPM = y \circ C_1 \circ C_2 \circ C_3 wo \circ, y = \text{ي}, wo = \text{و}$
 $PrPF = T \circ C_1 \circ C_2 \circ C_3 N \circ, T = \text{ت}, N = \text{ن}$
 $\}$
 $:$
 $:$
 $:$
 $:$
 $:$
Adjective 6: $C_1 \circ C_2 V_2 C_3 V_3 a \circ n, a = \text{ا}, n = \text{ن}$
Adjective 7: $C_1 V_1 C_2 \circ y C_3 \circ a \circ, a = \text{ا}$
Exaggeration formula 1: $C_1 \circ C_2 \circ a C_3, a = \text{ا}$
Exaggeration formula 2: $m \circ C_1 \circ C_2 \circ a C_3, a = \text{ا}$
Exaggeration formula 3: $C_1 \circ C_2 \circ wo C_3, wo = \text{و}$
Exaggeration formula 4: $C_1 \circ C_2 \circ y C_3, y = \text{ي}$
Exaggeration formula 5: $C_1 \circ C_2 \circ C_3$
Exaggeration formula 6: $m \circ C_1 \circ C_2 \circ C_3$
Exaggeration formula 7: $m \circ C_1 \circ C_2 y C_3, y = \text{ي}$
Name of preference: $a \circ C_1 V_1 C_2 V_2 C_3 V_3, a = \text{ا}$
Name of place and time: $m \circ C_1 \circ C_2 \circ C_3, m = \text{م}$
instrument name 1: $m \circ C_1 \circ C_2 \circ C_3, m = \text{م}$
instrument name 2: $m \circ C_1 \circ C_2 \circ a \circ C_3, m = \text{م}, a = \text{ا}$
instrument name 3: $m \circ C_1 \circ C_2 \circ C_3 haa, m = \text{م}, haa = \text{ه}$
 $\}$

Where , *PrSM:* Present Singular Male, *PrSF:* Present Singular Female, *PrDM:* Present Dual Male, *PrDF:* Present Dual Female, *PrPM:* Present Plural Male, *PrPF:* Present Plural Female, *PrSM:* Present Singular Male, *PrSF:* Present Singular Female, *PrDM:* Present Dual Male, *PrDF:* Present Dual Female, *PrPM:* Present Plural Male, *PrPF:* Present Plural Female, *PrSM:* Present Singular Male, *PrSF:* Present Singular Female, *PrDM:* Present Dual Male, *PrPM:* Present Plural Male, *PrPF:* Present Plural Female, *PSM:* Past Singular Male, *PSM:* Past Singular Female, *PD:* Past Dual, *PPM:* Past plural Male, *PPM:* Past plural Female, *ISM:* Imperative Singular Male, *ISF:* Imperative Singular Female, *ISF:* Imperative Dual, *IPM:* Imperative plural Male, *IPF:* Imperative plural Female, *OSM:*

Object Single Male, OSF: Object Single Female, *OSM:* Object Single Male, *OSM:* Object Single Male, *OSM:* Object plural Male, *OSM:* Object plural Female, *ASM:* Actor Single Male, *ASF:* Actor Single Male, *ADM:* Actor Daul Male, *ADF:* Actor Daul Femal, *APM:* Actor Plural Male, *APF:* Actor plural Femal, *PrSM:* Present Singular Male.

V.MED CLASIFICATION METHOD

Classification methods based on ME classification specificity assume that it is possible to determine the importance of a sentence on the basis of the words which constitute it. The most common way to achieve this is to find all drevative morphology in the document, and find the sentences which contain these words. A classification of the source is produced by extracting the words which words and its derivatives that appearance in the other document.

In this section we present the classification algorithm which depend on collect words which contains words and derivation.

Algorithm 3: Maximum Entropy Derivative Algorithm

Input: a set of Drivatives Words,extracted by ME

FAD = { $w_{1D}, w_{2D}, \dots, w_{nD}$ }. where, w_{iD} is a set of derivatives for the word $w_i, 1 \leq i \leq n$:

$*/w_{iD} = \{w_{iD_1}, w_{iD_2}, \dots, w_{iD_m}\}$

Output: classification documents

Method

$\forall w_{iD_j} \in w_{iD}, 1 \leq i \leq n, 1 \leq j \leq n$

If $w_{iD_j} \in DW, DW$ is the document words

$\forall w_{iD_j}$ assign the documents to the class which contain w_{iD_j}

End

VI.CONCLUSION AND FUTURE WORK

Arabic morphology, though considerably more difficult than the morphology found in the commonly studied European languages is fully susceptible to derivative analysis techniques. This paper presented an adaptation of existing Arabic morphological analysis techniques to make them suitable for the requirements of AIR applications. In this paper, derivation frames based on morphological and knowledge bases of verb lexicons can be related to produce a detailed representation of texts. The technique described in this paper has been applied to classification . For the future work an experiment in large Arabic data will be developed.

Reference

- [1] Atlam, E.-S., Morita, K., Fuketa, M. & Aoe, J. "A New Method for Selecting English Compound Terms and its Knowledge Representation". Information Processing Management Journal, 38 (6), 807-821, (2002).
- [2] Atlam, E.-S., Fuketa, M., Morita, K. & Aoe, J. "Document Similarity measurement using Field association terms". Information Processing & Management Journal, 39 (6), 809-824, (2003).
- [3] Atlam, E.-S., G. Elmarhomy, Fuketa, M., Morita, K. & Aoe, J. "Automatic Building of New Field Association Word

- Candidates Using Search Engine". Information Processing & Management Journal, 42 (4), 951-962, (2006).
- [4] Antworth, E. PC-KIMMO: a two-level processor for morphological analysis. In Occasional Publications in Academic Computing. Dallas, TX: Summer Institute of Linguistics. (1990).
- [5] Al-Kharashi, I. & Evens, M. "Comparing Words, Stems, and Roots as Index Terms in an Arabic Information Retrieval System". Journal of the American Society for Information Science and Technology, 45 (8), 548-560, (1994).
- [6] Sami Boudelaa, Friedemann Pulvermüller, Olaf Hauk, Yury Shtyrov, William Marslen-Wilson, Arabic Morphology in the Neural Language System, Journal of Cognitive Neuroscience, Vol. 22, No. 5, Pages 998-1010: 998-1010. 2010
- [7] Comrie, B. "The world's major languages" London-Sydney: Croom Helm, 1987.
- [8] Ezra Daya, Dan Roth, Shuly Wintner .Unsupervised Learning of Morphology, Computational Linguistics, Vol. 34, No. 3, Pages 429-448: 429-448. 2008
- [9] Fuketa, M., Lee, S., Tsuji, T., Okada, M., & Aoe, J. I. "A document classification method by using field association words". Information Science, 126, 57-70. (2000).
- [10] Halteren H. V., "New feature Sets for Summarization by Sentence Extraction," IEEE Intelligent Systems, vol. 18, no. 4, pp. 34 - 42 ISSN: 1541-1672. Doi:10.1109/MIS.2003.1217626, 2003.
- [11] Hmeidi, I., Kanaan, G. & Evens, M. Design and Implementation of Automatic Indexing for Information Retrieval with Arabic Documents. Journal of the American Society for Information Science and Technology, 48 (10), 867-881. (1997).
- [12] Ali Idrissi, Jean-François Prunet, Renée Béland, On the Mental Representation of Arabic Roots, Linguistic Inquiry Spring, Vol. 39, No. 2, Pages 221-259: 221-259. 2008
- [13] John J. McCarthy. A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry*, 12(3):373-418. 1981.
- [14] John J. McCarthy." formal problems in Semitic phonology and morphology" PhD thesis , Harvard College ,1975
- [15] Jing and Hongyan. Sentence reduction for automatic text summarization. In Proc. Conference on applied natural language processing, pp. 310-315. 2000.
- [16] Kiraz, G.. Arabic Computational Morphology in the West. In Proceedings of The 6th International Conference and Exhibition on Multilingual Computing, Cambridge. (1998)
- [17] Karlsson, F. "General linguistics" Helsinki: Helsinki university press, 1998.
- [18] J. Kupiec , J. Pedersen & F. Chen, A trainable document Summarizer. In Proceedings of annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'95) (pp. 68-73). (1995)
- [19] George Kiraz, Identifying Semitic Roots: Machine Learning with Linguistic Constraints, Computational Linguistics September 2008, Vol. 34, No. 3, Pages 459-462: 459-462.
- [20] Mani L., "Recent Developments in Text summarization' Proceedings of the tenth international conference on Information and knowledge management CIKM'01, 2001, pp. 529-53 ISBN:1-58113-436-3 doi:10.1145/502585.502677
- [21] Matthews, P.H. "Morphology" Cambridge university press, 1991.
- [22] Ryan McDonald, Joakim Nivre, Analyzing and Integrating Dependency Parsers , Computational Linguistics March 2011, Vol. 37, No. 1, Pages 197-230: 197-230.
- [23] Tsuji, T., Nigazawa, H., Okada, M. and Aoe, J. Early Field Recognition by Using Field Association Words". In the Proceeding of the 18th International Conference on Computer Processing of Oriental Language, 2, (1999), 301-304.
- [24] Tsuji, T., Fuketa, M., Morita, K. & Aoe, J.. An Efficient Method of Determining Field Association Terms of Compound Words. Journal of Natural Language Processing. 7(2), 3-26. (2000).
- [25] Abdelhadi Soudi, Antal van den Bosch, and Günter Neumann, Arabic Computational Morphology, ISBN 978-1-4020-6046-5 (e-book), Published by Springer, 2007
- [26] C. D. Paice. Method for evaluation of stemming algorithms based on error counting. Journal of the American Society for Information Science, 47 (8): 632 649, 1996.
- [27] C. D. Paice, Constructing literature abstracts by computer: techniques and prospects. Information Processing and Management, 26, 171-186. (1990)
- [28] K. Zechner, Fast generation of abstracts from general relevant sentences. In Proceedings of the 16th international conference on computational linguistic (COLING'96) (pp. 986-989). (1996).