British Journal of Mathematics & Computer Science
14(3): 1-9, 2016, Article no.BJMCS.23055
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Using Arabic Skeleton Morphology and Maximum Entropy for Arabic Document Classification

O. G. El-Barbary[1,2*]

[1]*Faculty of Science, Tanta University, Egypt.*
[2]*Faculty of Science and Arts in Sajir, Shaqra University, KSA.*

Author's contribution

The sole author designed, analyzed and interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/BJMCS/2016/23055
Editor(s):
(1) Victor Carvalho, Polytechnic Institute of Cávado and Ave, Portuguese Catholic University and Lusiada University, Portugal.
(2) Tian-Xiao He, Department of Mathematics and Computer Science, Illinois Wesleyan University, USA.
Reviewers:
(1) P. Moodley, University of Pretoria, South Africa.
(2) Solehah Yaacob, International Islamic University, Kuala Lumpur, Malaysia.
(3) Anonymous, Badji Mokhtar University, Algeria.
Complete Peer review History: http://sciencedomain.org/review-history/13194

Original Research Article

Received: 12th November 2015
Accepted: 26th January 2016
Published: 6th February 2016

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 maximum entropy and morphological derivation of 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.

1 Introduction

Document Classification (DC) is the process of grouping similar document in the same class or field. DC has rapidly grown into a major research area [1-13] as illustrated by the Document Understanding Conference (DUC) and Text Analysis Conference (TAC) series. Arabic classification systems are still not as

**Corresponding author: E-mail: omniaelbarbary@yahoo.com, ualbarbari@su.edu.sa;*

sophisticated and as reliable as those developed for other languages like English. This is due to the richness and complexity 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.

Arabic is one of the most widely used languages in the world, yet there are relatively few studies on retrieval and classification of Arabic documents. Arabic is one of the six official languages of the United Nations. It is the mother tongue of 300 million people. Developing text classification for Arabic documents is a challenging task due to the complex and rich nature of the Arabic language. Classifying Arabic text is different than classifying English language because Arabic is highly inflectional and derivational language which makes monophonic analysis a very complex task. The Arabic language consists of 28 letters, is written from right to left and the majority of words have a tri-letter root. It needs to be stated that the N-gram as an indexing system is used to classify documents but it is still insufficient to classify Arabic documents. Therefore, this study proposes an innovative means to classify documents by using maximum entropy and Arabic morphological derivatives.

The Arabic morphology considers special challenge to computational natural language processing system .Arabic is a complicate languge due to its morphology, this reason produces Arabic IR researches depends on Arabic morphology [14-21]. Arabic morphology becomes an vital part of many Arabic information retrieval system. Arabic offers unusual challenges for data driven. any word in Arabic language is 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. Many research in AIR based on stem [22,23] stemming is a method used in IR. 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, a new technique for Arabic document classification is devolped using maximum entropy and Arabic morphological derivatives. To discover the derivatives for words, we 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 derivatives frame 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 "كَاتب, kaatab which means in English writer" as Active Participle.

The rest of this paper is organized as follows; Section 2 defines maximum entropy and how use it to document classification. In Section 3 Arabic morphology was discussed. Section 4 our new algorithm to extract all derivatives of FA words will illustrate. Section 5, explains Conclusion and future work.

2 Maximum Entropy for Text Classification

2.1 Maximum entropy modeling

Maximum entropy is a technique for learning probability distribution from data. Maximum Entropy was in the beginning used to guesstimate how much of the data can be packed together before they are transmitted over a communication channel [24,25,26]. Entropy measures the average uncertainty of a single random variable X:

$$\mathrm{H(p)} = \mathrm{H(X)} = \sum_{x \in X} \mathrm{P(x)} \log_2 \mathrm{P(x)},$$

Where p(x) is the probability mass function of the random variable X.

On the other hand, in natural language processing, we want to locate a model to make the most of the entropy.

Numerous troubles in natural language processing can be re-formulated as statistical classification problems. particularly, in text classification we consider the text classification task to be a process W which takes as input a document d and produces as output a class c. The output of the process W may be affected by some contextual information X, whose domain is all the possible textual information contained in the document d.

2.2 Maximum entropy for text classification

The rule of Maximum Entropy (ME) is base the idea that when estimate the possibility distribution, one should select that distribution which leaves you the biggest residual 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 for 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:

$$f_{w,\grave{c}}(d,c) = \begin{cases} 0 \quad if \;\; c \neq \grave{c} \\ \dfrac{N(d,w)}{N(d)} \quad otherwise, \end{cases}$$

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 natural binary features. In text classification, we expect that features accounting for the number of times a word occurs should improve classification.

3 Arabic Morphology

Morphology is Part of linguistics which study word formation and structure. It is consists of inflectional morphology and derivational morphology [27,28,29,30]. Inflectional morphology is distinct as the apply of morphological methods to form, inflected word forms from a lexeme. Inflection word forms show grammatical associations between words. However, derivational morphology is apprehensive with the derivation of new words from other words using derivational affixes. A root or a morpheme is the minimum item of language which has a meaning [31,32]. The morphological handing out is needed in natural language processing to handle inflected word forms. An Arabic word composed of a stem. A stem is consists of a consonantal root and a pattern. also, it contains affixes and vowels. There are 15 trilateral form, of which at least 9 are familiar. Within each conjugation pattern, an entire pattern is found. Two voices (active and passive) and tense (perfect and passive) and five moods (indicative, subjective, jussive, imperative and energetic).

3.1 The Arabic vowels

Arabic has 3 vowels,َ Fat-ha (A), ِ Kasra (E/I) 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 (ـِيْ Yaa (EE) , ـا Aleph (AA), ـُوْ Waw (UU), ـَيْYaa Leen (EI), ـَوْ 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 (مَـٰلِكَ: maaleka, which means owner in English) and the word (مُلْكَ: moleka which means Sovereignty in English).Also, the word (ميِّت: myaat which means the person who will die but he still alive in english) and the word (ميْت: myt which means the person who actually died in English). Also, the word (كَره : karah which means in English the hardship which affect rights outside of itself and with force) but the word (كُره : 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.

4 Derivatives Frame and Its Algorithm

Arabic morphology contains 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 "فُتُح" which means opening in English. The important derivatives as follows:

1- Actor the person who opens, "فاتح" means opener.
2- Object, Signed by the act like "مَفتوح" means the opened.
3- The place where opening "مَفتح".
4- Machine name or the instrument of opening "مُفتاح", 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 1: FAD word algorithm

Input: *a set of words*

Output: *a set of DW sets, DW= {* $w_{1D}, w_{2D}, \dots, w_{nD}\}.$ *where,* w_{iD} *is a set of derivatives for the word* $w_i, 1 \leq i \leq n.$ $:*/w_{iD}$ *= {* $w_{iD1_1}, w_{iD_2}, \dots, w_{iD_m}$*}*

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*=* $C_1C_2C_3$ *,where C's is an actual letter or*

then

{

PrSM= yC_1 C_2 C_3*, y=*ي

PrSF= TC_1 C_2 C_3*, T=*ت

PrDM= y C_1 C_2 C_3*, y=*ي *, a=*ا

PrDF= TC_1 C_2 C_3*,T=*ت *, a=*ا

PrPM= yC_1 C_2 C_3*, y=*ي *, wo=*وا

PrPF= T C_1 C_2 C_3*, T=*ت *, N=* ن

}

If (V_1 = ◌َ ∧ V_2 = ◌َ*)* then

{

{

PrSM= y ◌َC_1◌ْC_2◌َC_3◌ُ*, y=*ي

PrSF= T ◌َC_1◌ْC_2◌َC_3◌ُ*, T=*ت

PrDM= y ◌َC_1◌ْC_2◌َC_3*a* ◌َ*, y=*ي *, a=*ا

PrDF= T ◌َC_1◌ْC_2◌َC_3*a* ◌َ*, T=*ت *, a=*ا

PrPM= y ◌َC_1◌ْC_2◌َC_3*wo* ◌ُ*, y=*ي *, wo=*وا

PrPF= T ◌َC_1◌ْC_2◌َC_3*N* ◌َّ*, T=*ت *, N=* ن

}

Or

{

PrSM= y ◌َC_1◌ْC_2◌ِC_3◌ُ*, y=*ي

PrSF= T ◌َC_1◌ْC_2◌ِC_3◌ُ*, T=*ت

PrDM= y ◌َC_1◌ْC_2◌ِC_3*a* ◌َ*, y=*ي *, a=*ا

PrDF= T ◌َC_1◌ْC_2◌ِC_3*a* ◌َ*, T=*ت *, a=*ا

PrPM= y ◌َC_1◌ْC_2◌ِC_3*wo* ◌ُ*, y=*ي *, wo=*وا

PrPF= T ◌َC_1◌ْC_2◌ِC_3*N* ◌َّ*, T=*ت *, N=* ن

}

Or

{

PrSM= y ◌َC_1◌ْC_2◌ُC_3◌ُ, y=ي

PrSF= T ◌َC_1◌ْC_2◌ُC_3◌ُ, T=ت

PrDM= y ◌َC_1◌ْC_2◌ُC_3a ◌َ, y=ي , a=ا

PrDF= T ◌َC_1◌ْC_2◌ُC_3a ◌َ, T=ت , a=ا

PrPM= y ◌َC_1◌ْC_2◌ُC_3wo ◌ُ, y=ي , wo=وا

PrPF= T ◌َC_1◌ْC_2◌ُC_3N ◌َّ, T=ت , N= ن

}

:

:

:

:

Adjective 6: C_1◌ُ◌ْ$C_2V_2C_3V_3a$◌َn , a = ا, n = ن

Adjective 7: $C_1V_1C_2$◌ْyC_3◌َ a◌َ , a = ى

Exaggeration formula 1: C_1◌َC_2◌َّ a C_3 , a = ا

Exaggeration formula 2: m◌ِC_1◌ْC_2◌َ a C_3 , a = ا

Exaggeration formula 3: C_1◌َC_2◌ُ wo C_3 , wo = و

Exaggeration formula 4: C_1◌َC_2◌ِ y C_3 , y = يـ

Exaggeration formula 5: C_1◌َC_2◌ِ C_3

Exaggeration formula 6: m◌ِC_1◌ِC_2◌ْ C_3

Exaggeration formula 7: m◌ِ C_1◌ْC_2y C_3 , y = يـ

Name of preference: a◌َ $C_1V_1C_2V_2C_3V_3$,a=أ

Name of place and time: m◌َ C_1◌ْC_2◌َC_3 ,m= مـ

instrument name 1: m◌ِ C_1◌ْC_2◌َC_3 ,m=مـ

instrument name 2: m◌ِ C_1◌ْC_2◌َ a◌َC_3 ,m=مـ , a=ا

instrument name 3: m◌ِ C_1◌ْC_2◌َC_3haa ,m=مـ , haa=ة

}

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.

5 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 derivative 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 depends on collects words which contains words and derivation.

Maximum Entropy derivative Algorithm takes the input as a set of derivative words, which every w_i is belongs to a class. For every $w_{iD_j} \in w_{iD}$ the documents will assign to the class which contain the words.

Algorithm 2: Maximum entropy derivative algorithm

Input: *a set of Drivatives Words, extracted by ME*

FAD= $\{ w_{1D}, w_{2D}, \ldots, w_{nD}\}. where, w_{iD}$ *is a set of derivatives for the word* $w_i, 1 \leq i \leq n.$ **:**
$*/w_{iD} = \{ w_{iD_1}, w_{iD_2}, \ldots, w_{iD_m}\}$

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

6 Conclusion and Future Work

The paper offered an alteration of existing Arabic morphological analysis technique to make them fit 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.

Competing Interests

Author has declared that no competing interests exist.

References

[1] Atlam ES, Morita K, Fuketa M, Aoe J. A new method for selecting English compound terms and its knowledge representation. Information Processing Management Journal. 2002;38(6):807-821.

[2] Atlam ES, Fuketa M, Morita K, Aoe J. Document similarity measurement using field association terms. Information Processing & Management Journal. 2003;39(6):809-824.

[3] Atlam ES, Elmarhomy G, Fuketa M, Morita K, Aoe J. Automatic building of new field association word candidates using search engine. Information Processing & Management Journal. 2006;42(4): 951-962.

[4] El-Halees A. A comparative study on Arabic text classification. Egyptian Computer Science Journal. 2009;20:110-115.

[5] The Islamic University Journal (Series of Natural Studies and Engineering). Arabic Text Classification Using Maximum Entropy. Alaa M. El-Halees. 2007;15(1):157-167.

[6] Khorsheed MS, Al-Thubaity AO. Comparative evaluation of text classification techniques using a large diverse Arabic dataset. Language Resources and Evaluation. 2013;47:513-538.

[7] Meryeme Hadni, Said Alaoui Ouatik, Abdel Monaime Lachkar. Effective Arabic stemming based hybrid approach for Arabic text categorization. International Journal of Data Mining & Knowledge. Management. 2013;3(4):1.

[8] Halteren HV. New feature sets for summarization by sentence extraction. IEEE Intelligent Systems. 2003;18(4):34 – 42. ISSN: 1541-1672.
DOI: 10.1109/MIS.2003.1217626

[9] Jing and Hongyan. Sentence reduction for automatic text summarization. In Proc. Conference on Applied Natural Language Processing. 2000;310–315.

[10] Kupiec J, Pedersen J, Chen F. A trainable document summarizer. In Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95). 1995;68–73.

[11] Mani L. Recent developments in text summarization. Proceedings of the Tenth International Conference on Information and Knowledge Management CIKM'OI. 2001;529-53. ISBN:1-58113-436-3.
DOI: 10.1145/502585.502677

[12] Paice CD. Constructing literature abstracts by computer: Techniques and prospects. Information Processing and Management. 1990;26:171–186.

[13] Zechner K. Fast generation of abstracts from general relevant sentences. In Proceedings of the 16th International Conference on Computational Linguistic (COLING'96). 1996;986–989.

[14] Sami Boudelaa, Friedemann Pulvermüller, Olaf Hauk, Yury Shtyrov, William Marslen-Wilson. Arabic morphology in the neural language system. Journal of Cognitive Neuroscience. 2010;22(5):998-1010.

[15] Ezra Daya, Dan Roth, Shuly Wintner. Unsupervised learning of morphology. Computational Linguistics. 2008;34(3):429-448.

[16] 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. 1997;48(10):867–881.

[17] Ali Idrissi, Jean-François Prunet, Renée Béland. On the mental representation of Arabic roots. Linguistic Inquiry Spring. 2008;39(2):221-259.

[18] Kiraz G. Arabic computational morphology in the West. In proceedings of the 6th International Conference and Exhibition on Multilingual Computing, Cambridge; 1998.

[19] George Kiraz. Identifying Semitic roots: Machine learning with linguistic constraints. Computational Linguistics. 2008;34(3):459-462.

[20] Ryan McDonald, Joakim Nivre. Analyzing and integrating dependency parsers. Computational Linguistics. 2011;37(1):197-230.

[21] Abdelhadi Soudi, Antal van den Bosch, Günter Neumann. Arabic computational morphology. ISBN 978-1-4020-6046-5 (e-book), Published by Springer; 2007.

[22] 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. 1994;45(8):548–560.

[23] Paice CD. Method for evaluation of stemming algorithms based on error counting. Journal of the American Society for Information Science. 1996;47(8):632 649.

[24] Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification; 1999.

[25] Stanley F. Chen, Rosenfeld R. A Gaussian prior for smoothing maximum entropy models. Technical report CMU-CS-99108, Carnegie Mellon University; 1999.

[26] Ronald Rosenfeld. Adaptive statistical language modelling: A maximum entropy approach. PhD Thesis, Carnegie Mellon University; 1994.

[27] Antworth E. PC-KIMMO: A two-level processor for morphological analysis. In Occasional Publications in Academic Computing. Dallas, TX: Summer Institute of Linguistics; 1990.

[28] Comrie B. The world's major languages. London-Sydney: Croom Helm; 1987.

[29] Karlsson F. General linguistics. Helsinki: Helsinki University Press; 1998.

[30] Matthews PH. Morphology. Cambridge University Press; 1991.

[31] John J. McCarthy. A prosodic theory of nonconcatenative morphology. Linguistic Inquiry. 1981;12(3):373-418.

[32] John J. McCarthy. Formal problems in Semitic phonology and morphology. PhD Thesis, Harvard College; 1975.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://sciencedomain.org/review-history/13194