Arabic Minimal Pairs Word Detection and Disambiguation

Mohamed Taybe Elhadi
Software Engineering Department, Faculty of Information Technology,
Az Zawiyah University, Az Zawiyah Libya

ABSTRACT

This work is an attempt to solve a common writing problem and pitfall for Arabic language. The problem involves words that contain letters such as (among many others) ظ THA and ض DHA. The problem involves terms that are formally minimal pairs (more precisely near minimal pairs), near homographs (homophones), it requires determination of the right term and resolutions of created ambiguities. It is not just embarrassing to the authors, but in many situations, it results in wrong usage of words and consequently can lead to an ambiguous sentence(s). It becomes difficult to interpret such words or sentences, especially by computer involved in applications such as information retrieval, language translations and summarizations. A very amalgamated determination process was suggested that is comprised of multiple stages of feature selection, classifier selection and classification. A sample set of terms selected with a reasonable success rate making classifiers accuracies vary, but overall, all terms are reasonably accurate and close in values. MCC values are also variable with some reasonable good ranges. It is notable that some classifiers did not converge and the MCC is set to zero. Considering results obtained from classifiers with highest training rates and those with highest MCC, It can be easily concluded that Random Forest algorithm is the champion with high accuracy in most of the terms, and many times very close to the highest rate, classifiers were close. It also scored the highest for the mean values calculation across all terms. We can easily say that a combination of extracted features from a corpus along with machine learning classification techniques, the problem can be solved with high accuracy.

General Terms

Natural language processing, Machine learning, classifications

Keywords

Minimal Pairs, Disambiguation, Machine Learning, Classifiers, Arabic Language.

1. INTRODUCTION

A minimal pair is a term or word that differs from another paired word by a single phoneme (sound unit). In English, for example, the word **tip** paired with the word **sip** is a minimal pair. In Arabic there are many minimal pairs that contribute to ambiguities of words causing anomalies in writing, pronunciation and word usage in general. For example, the word (ضن) and its close relative (ظن) differ only in first letter (related in graphing and to some extent in phoneme), but mean totally different things. The first means doubtfulness while the second means stinginess. In some situations, paired words may have opposite meanings for example to speak highly of someone (تقريظ)and to speak negatively of someone(تقريض) but differ only in one character.

Even though, minimal pairs are normally looked at from the sound perspective, the word pairs addressed here, are seen as a cause of ambiguities from a written perspective. In doing so one would like to consider this as a "word pairs disambiguation" problem rather than a minimal pairs problem.

In Arabic, ambiguity is the love to hate things. Many linguists argue for it while many others equally argue against it. In natural language in general and in Arabic in particular from the point of view of some literature experts like poets and religious transcriptionists such vagueness is inevitable or rather inevitable at times. Some types or levels of ambiguities are considered as a remarkable feature of language production. They provide for flexibility in language use and interpretations giving such users more expressiveness and flexibility. Not the case, however when it comes to computers. Machines and algorithms need to remove or reduce ambiguities in order to involve in any meaningful natural language processing endeavor.

Ambiguities can range from difficulty in interpreting the exact and right meaning (sense) for single word to more complicated ambiguities caused by full sentenced or even more complicated language constructs. Word Sense Disambiguation (WSD) is the basic and most common disambiguation task concerned with ability to resolve ambiguities or vagueness in interpreting words by assigning a particular and certainly the most sensible sense to an ambiguous word as intended by the producers. A number of categories or sources of word ambiguities are presented in literature including homograph, homophone or homonym …etc.

In most languages like English, a homograph is when two words have different spelling, sound and meaning like the word "**lead**" (to go in front of) and "**lead**" (a metal) or "**wind**" (to follow a course that is not straight) and "**wind**" (a gust of air).

A homophone, however, is the case of more than a word with thesame sound but a different meaning. Homophones may or may not have the same spelling asin the example of "**to/two/too**" and "**pray/prey**". A homonym is when a word is spelled similar to another but has a different sound and meaning, or a word that sounds like another but has a different spelling and meaning (homophone) or both. That is to say a word that is spelled and pronounced like another but has a different meaning [1].

A particular situation in Arabic language that relates to minimal pairs as well as to word pairs ambiguity causing difficulties from the perspective of authors by inducing uncertainty as to what is the right word to use. A very common scenario of confusion is to do with words containing the letters (among many others) ظ THA (referred to as Z-

Words from now on) vs. those that contain the letter ض DHA (referred to as D-Words from now on). DHA (ض) is a letter that is many times referred to as the distinguishing letter for Arabic language and that is why Arabic is known as the language of DHA. It is estimated that Arabic has 900 terms that contain the letter DHA. On the other hand, there is a set of terms that are very similar in form but differ in sounds and meanings. Arabic contains approximately 90 plus Z-words, about a third (32 words) of which are in common use. Table 1 contains a list of the most commonly used words.

The issue can be summarized as a general confusion in word meaning and usage when it comes to words containing either letter THA or DHA. In particular, there are words that spell the same except for the letters THA and DHA (ظ and ض). Table 2 contains some examples of words that differ in meaning and in some cases they may have opposite meanings.

Table 1. Sample list of common THA/DHA words with associated meanings

DHA - Words		THA – Words	
العَظَل: وهو الشدة، من: أمر معظل. الغيظ: الحنق، سخط، غضب شديد. الفظاظة: القسوة. الفظاعة: من الأمر الشنيع. التقريظ: مدح الحي بالشعر. المواظبة: الاستمرار. الكظم: كتم الحزن. اللحظ: النظر.	اللفظ: مفردة. النّظْمُ: الأنظمة. النظافة: طهارة. النظَر: الشوف العظم: ما بني عليه اللحم. العظيم:صفة واسم من أسماء الله الوظيفة: العمل. اليقظة: ضد النوم.	الظعن: السفر بالنساء. الظرف: الحال أو الوعاء. الظريف: البارع الفكاهي الممتع الظِّنّ: الشك. الظِّلّ: الخيال، كنف وحماية. الظفر: ضد الخيبة. الظهر: وقت، من أجزاء الجسم الظماء: العطش.	الحَظّ:النصيب. الحِفْظُ: ضد النسيان. الحَظْرُ: المنع. الحَظْوَةُ: الرفعة. الظلم: الجور. الظليم: ذكر النعام. الظبي: الغزال. الظبة: طرف السيف.

With people's inability to distinguish the sound and the use of such words, the problem can be considered like near homophones (even though they are not but very close to homophones).

Table 2: Sample words differing in the letter D or Z with different/opposite meanings

The Z-Word: D-Word	Z-Meaning D-meaning	The Z-Word: D-Word	Z-Meaning D-meaning
الحظ :الحض	Luck/Motivating	ظاف: ضاف	Add/Host
حظر: حضر	Prohibition/ Present	نظر: نضر	Sight/Good looks
ظل: ضل	Shadow/ Go astray	نظرة: نضرة	A Look/Good looking
ظن : ضن	Doubtfulness/ Stinginess	نظير : نضير	Equivalent/good looking
فاظ: فاض	Overflow/ Die	ناظرة ناضر:	Looking at/looking good
الظلع: الضلع	Side/Handy Cabbing	الغيظ: الغيض	Temper/Little

The problem of such word pairs is not limited to Z and D words but many others. The letters (س) sah and (ص) sauh as in (سلاح/صلاح) (weapon/goodness), (ت) ta and (ث) tha in some Arabic communities as in (تابت و ثابت) (repented/fixed) and many others. We are focusing on Z-D terms to highlight the issue and are considered the most difficult and cause much more embarrassment with both writing and speaking. What is done, however, applies to all. A set was selected as a sample to use. It was taken from literature and analysis done of some available Arabic corpora [2]. Different surveys express similar sets with slight variations. Table 1 shows 32 words normally considered as an active set of words. Table 2 shows the selected to be analyzed words.

2. RELATED WORK

This work wasbased on a number of fields and relates to a number of research issues. For completeness, a brief mention of such fields and issues was made. WDS and its related fields are relatively young research areas and aremostly done for English language. Recently though, interest has spread to other word languages including Arabic. WSD is a way to find the right sense of an ambiguous word in a particular context. It is normally looked at and classified from a number of aspects. It can be classified by the words used (sample vs complete),or based on sources of knowledge and type of reasoning methods (Knowledge-based, Supervised and Unsupervised) [3-7]

2.1 Minimal Pair Words

A minimal pair is a word that differs from another paired word by a single phoneme. The two phonemes in the minimal pair words tip and sip (- /t/ and /s/ -) are considered minimal and called minimal pairs in reference to least possible, that is least possible distance characterizes minimal pairs [8-12].The phonemes /s/ and /t/ have few contrastive differences in relation to place, manner and voice.

2.2 Word Sense Disambiguation (WSD)

Importance of WSD stems from the many applications it relates to, including Machine Translation [13,14], Information Retrieval [15,16], Text-Data Mining, Lexicography [17], and Information Extraction [18]. WSD is an important requirement of machine translation because of that fact that word translations are context dependent. Resolving ambiguity in a query is the most vital issue in Information Retrieval systems. It also plays an important role ininformation extraction and summarization and in many other research field suchas Bioinformatics, Named Entity recognition, system, co-reference resolution.

WSD is the selection and assignment of a particular sense to a specific word based on contextual and non-contextual evidences. It can be seen as a classification process with word senses making the classes, and the context containing the evidence. Occurrence of a particular word is given one or more of its possible sense classes using available evidence. The correct meaning of an ambiguous word is determined by its context and other worldly knowledge. In case of a human, such ambiguities can be resolved without much difficulty. Computationally, however, is not an easy task but a difficult and an AI-complete problem [19].

Its difficulty stems from several factors including but not limited to: (1) strong reliance on knowledge ranging from collections of texts, to more structured resources, like dictionaries, thesauruses, semantic networks, and ontologies [20]; (2) Richer lexical knowledge resources and large-scale coarse-grained sense inventories, are becoming available to help in WSD.

Principally, WSD can fall into one of the following types: (1) Lexical samples where the requirement is to disambiguate a restricted set of target words. This is mostly handled by Supervised systems trained using a number of hand-labelled instances; and (2) All-words WSD, where all open-class words such as nouns, verbs, adjectives, and adverbs are to be disambiguated. More in this as it relates to Arabic is discussed in next sections.

2.3 Arabic Language

The Arabic language has been receiving more attention by the world in general and by the scientific communities in particular. It is the main language in the Arab world and the secondary language in many other countries. Arabic is the official language of 25 countries and is spoken by over 250 million just in the Arab world alone and by close to 400 million all over the world. The US Department of Cultural Affairs categorized Arabic, among other world languages, as a critical language. The United Nations heavily emphasizes the social and political importance of these languages. It lists Arabic as one of the six official languages of the United Nations [21]. Furthermore, more than 3% of the internet content is Arabic content, which puts it in the fourth rank.

Arabic Language consists of 28 alphabet characters written from right to left. Its letters have different styles when appearing in a word depending on the letter position. Arabic words have two genders, feminine and masculine; three numbers, singular, dual, and plural; and three grammatical cases, nominative, accusative, and genitive. Words are classified into three main parts of speech, nouns (including adjectives and adverbs), verbs, and particles. All verbs and some nouns are morphologically derived from a list of roots. Words are formed by following fixed patterns. The prefixes and suffixes are added to the word to indicate its number, gender, and tense [22].

The Arabic language has rich morphology and a complex orthography. Different shapes and diacritics of the Arabic language make pairing a difficult task due to this complex morphology. Limited access to technology has hindered research in automation and utilization of Arabic.

Arabic is a Semitic language known for its morphological roots and driven by its affixes. It is a derivational and inflectional language with 85% of its words derived from close to10.000 independent roots [23]. It has a relative free word order and several dialects in addition to Modern Standard Arabic (MSA) which is the standard in media, news channels and educational institutes.

2.4.1Arabic Word Sense Disambiguation

WSD is one of the oldest problems in computational linguistics [22]. It came up as part of machine translation in the 40s. It was resurrected in the 1970s within Natural Language Understanding in artificial intelligence research. In the 80s, largescale lexical resources and corpora became available with knowledge extracted automatically from the resources replacing old manual ways. Dictionary-based WSD had begun and the relationship of WSD to lexicography became explicit and Naïve Bayes with Decision Tree, Rule Based Learner, Probabilistic Model were alsoused in disambiguation [24].

A relatively complex nature of Arabic makes for many challenges for NLP and applications such as WSD. According to [25,26], Arabic difficulty for WSD is caused by multiple factors and sources including the following:

- Missing of diacritics: Existence of several possible meanings for non-vocalized Arabic in modern editions, where the texts are not vocalized. A word like (ktb , كتب) has sixteen vocalizations [25]. The word "ولد " (walad) like in ولد أَحمد مَنذ شَهر (Ahmed was borna month ago). Verb ولد هَذا اَلرّجل مَريض(This man's son is sick).
- Agglutinative nature: In order to understand the meaning of words, a prior segmentation is needed. An example would be (وتتذكروننا) Watatathakarounana which means "you remember us".
- Lack of language resources: Arabic language research badly lacks resources such as dictionaries, thesaurus, and previously tagged corpora [26]. Arabic language has seen limited research and publications on WSD due to the lack of resources for the Arabic language [27].

Like other languages, and as presented in [28] approaches classified using thissource of knowledge are based on dictionaries, thesaurus and are called knowledge-based methods. Alternatively, the unsupervised approach of Bootstrapping Arabic Sense Tagging [29], make use of translational equivalences to parallel Arabic English corpus for the annotation of Arabic text using English WordNet taxonomy. A 90% Accuracy was obtained.

In [30], a Naïve Bayes Classifier is trained on labeled corpus where decision rule is applied to choose a class. In [31] WSD algorithm based on Thematic Words of a given context was introduced to choose the appropriate sense of an ambiguous word.

In [25], information retrieval measures were used to estimate the similarity between the context of use for each sense of the ambiguous word and the original sentence. The measures were combined with the Lesk algorithm to develop a system for Arabic word sense disambiguation. Exact string-matching algorithm [32] namely String-Matching algorithm of Boyer and Moore was then used to find the occurrences of a string in a text, it was proposed to use some information retrieval measures with the Lesk algorithm and it achieveda rate of 73% [25].

In [33] a Context matching algorithm returns a semantic coherence score corresponding to the context of use that is semantically closest to the original sentence. This algorithm achieveda precision of 78%.

According to [33] most WSD works evaluated in Semeval 2007 showed that supervised methods achieve the best disambiguation quality of about 80% precision and recall for coarse-grained WSD. Example of such systems are [34] which were based on "bootstrap" and with a reported Accuracy of 90% and [30] in which the Naïve Bayes algorithm was applied with dictionary to collect10 training samples for each word for the testing phase. This work achieveda rate of precision of 73%. In [35], based on "bootstrap", the Rocchio Classifier was rated and compared to Naïve Bayesian classifier, the most frequent sense and the support vector machine using Arabic lexical samples. The Rocchio classifier achievedan overall Accuracy of 88% [21,27,33].

2.5 Machine Leaning and Classifiers

Machine learning algorithms have become an integral part of many data analyses especially classification and prediction. Approaches for machine learning are divided into three main categories: (1) Supervised learning which is used if the available data to be used for training has a labeled attribute and other data does not contain a label; (1) Unsupervised learning which has no labeled information, but the algorithms strive to discover any existing pattern in the data. (3) Deep learning which learns and improves using artificial neural networks with larger, sophisticated neural networks that aid in classification problems and its different applications such as language translation, and speech recognition [36]. Supervised learning algorithms are the subject of these experiments. The following is a brief count of the machine learning algorithms used in this work:

- Random Forest is an algorithm that is based on theclassification and regression model of decision trees. Thisallows problem solving [37] with “if this then that” conditions ultimately yielding a specific result. Random Forest mitigates this problem well. by training on different samples of the data, RF which is a collection of decision trees whose results are aggregated into one final result in order to limit overfitting without substantially increasing error levels.
- Artificial Neural Networks (ANN) uses the Multi-Layer Perceptron classifier which relies on an underlying Neural Network to perform the task of classification [38].
- Naive Bayes (NB) is a simple probability model that is based on Bayes’ theorem and strong (naïve) independence assumptions between attributes. by training on different samples of the data. Bayes’ based classifiers are fast and easy to implement but use the simplistic assumption of independence of predictors. It has many applications [39].
- K-Nearest Neighbor (KNN) isa supervised learning classifier, which uses proximity to make classifications or predictions based on the assumption that similar points can be found near one another [40].
- Logistic Regression (LR) is used for predicting the categorical dependent variable using a given set of independent variables. It predicts a binary outcome, based on prior observations of a data set has the ability to provide probabilities and classify new data using continuous and discrete datasets based the well-known sigmoid function [41].
- Support Victor Machines (SVM) is used more in classification with the aim of finding a hyperplane in an N-dimensional space representing the number of features that distinctly classifies the data points. Support vectors are data points that are closer to the hyperplane which are used to maximize the margin of the classifier [42].

3. SUGGESTED PROCEDURE

The intention was to create a useable set of features using certain context and general structural evidence extracted from Arabic corpora examples. The features set is to be used to solve the problem of which word best fits the current writing context (Z-word vs. D-word). Some of the produced features are weighted using TF-IDF for the set of context words in both original and stemmed form. The following is the list of features used and how similarity is defined:

- Simple word frequency is the frequency of occurrence of the word in context (**FRQ**).
- Word similarity calculations using an arbitrary 6 word set around the relevant word. Three words before and three after if available. The six context words for each sentence are grouped (clustered) into one single dictionary entry using bag of words approach in two different ways:
    - The words were used in their original form without any pre-processing and without any associated weights (OCT).
    - The words weresubjected to a stemming using stemmer from [43] and each term was weighed using tf-idf [44] (OIDF).

In both cases cosine function was used for similarity calculations.

- More features are included using part of speech (POS) of the collected sentences.
    - Each sentence was POS tagged using camel-tools tagger [45].
    - The list of tags for each sentence was then converted into a string of a single character per tag for easy comparisons using sequence matching algorithms.
    - A single POS string for each dictionary category was created by concatenating a selected set of tags around the original word tag.

The POS similarity was determined in two ways:

- Using Longest Common Sub-sequence algorithm (strsimpy) on the single string representing the set of sentences (**LCSS**).
- Using the POS tag of the highest percentage (**WPC**).

The suggested procedure is performed through a number of basic steps applied on some datasets as outlined next.

3.1 Base Dictionary Creation

This is an important step in which the main dictionary wascreated. Following are the sub-steps that take place in the creation of the dictionary.

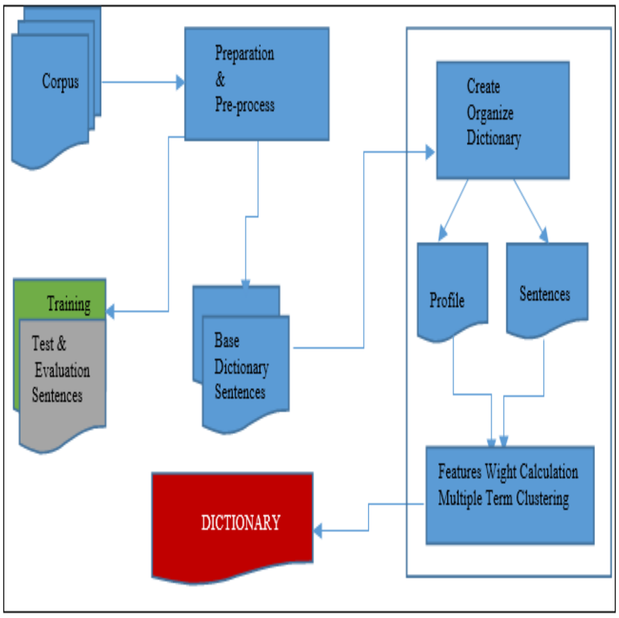
Corpus
Preparation & Pre-process
Create Organize Dictionary
Profile
Sentences
Training
Test & Evaluation Sentences
Base Dictionary Sentences
Features Wight Calculation Multiple Term Clustering
DICTIONARY

Figure 1: General Procedure-Dictionary Creation

3.1.1 Corpus Pre-processing and Preparation

Dividing the Corpus data into two data sets of sizable number of sentences. The first set compromised of 32,492 sentences is referred to as Base Dictionary Set and is used to create the base dictionary. The second set is made of 113,162 sentences and referred to as a Training and Evaluation set and is used for classifiers training and final validation.

This step involves a number of tasks as is shown next:

(1) Preparation of the text in which normalization and de diacritics removaltakes place.

(2) Selection of context words: define context, in this work was 3 positions before and 3after the relevant word were used.

(3) POS tagging and processing

(4) Dictionary Organization and Clustering: Dictionary items weregrouped together and the dictionary attributes weredivided into two sets:

- Dictionary Profile: Creation of Short profile which contains the frequency, contextwords before and after keyword, as well as POS information.
- Dictionary Sentences: Contains each entry or record that was created including sentence text and POS information.

(4) Finally, Dictionary Consolidation where single item or entry for each word category with relevant frequency of info wasmade into a single entry to optimize searching of the dictionary by reducing the size of dictionary.

Each sentence wasprocessed to produce the set of basic components or features (See Table 3)

Table 3: Item Profile (FEATURES & Sample Values)

okw:عضو **rkw**:عضو **ows**:اضيوعة اوضدء ا عضو عضو عضو عضو عضو اعضاء
www: خرر-0.41 طلق-0.41 كنو-0.41 للكترن-0.41 نهم-0.41 هجم-0.41
bkw: اي اي واقرت الدولي التي تضم الدولي التي تضم المكونه من دوله صعيد اخر قال عن الحكومه الالمانيه النادي قائم ب
akw: بقيمه مليون دولار في الاتحاد الاوروبي بالبرلمان العراقي ان مستقل فيه لمساعدتهم في
owp:noun noun noun noun noun noun noun

The result of this process wasa Dictionary containing intermediate data as shown in Table 3.The base dictionary wascreated using the selected corpus going through the following steps:

- Each file wasprocessed sentence by sentence selecting only those sentences that contain relevant words (those with THA or DAH in them).
- Profile Dictionary which, in addition to tty, id, fc, okw, rkw, containeditems based on the set of sentences that contain the relevant term.
- The new clustered dictionary (called Profile Dictionary) wascreated.

Table 4:ProfileDictionary Item(Features&Values**)**

Type: (tty:D)	**Unique id:** (id:4) **Frequency**(fc:1)
Root: (rkw:ضلل)	**Sentence:** (os: مكن ان يكون هذا الضل لجزء من جسم المركبه نفسها)
Original word: **POS**(w0p:noun_prop)	**SENT POS :** (Pos:verb conj_sub verb pron_dem noun_prop noun prep noun noun noun
SENT POS STR : (sPos:lalNhgfggg)	**context POS STR** : (nsPos:lalNhgfggg)

This new document is a higher-level of abstraction from the previous and is divided into two separate documents:

- Sentences Dictionary.
- Profile Dictionary.

Table 4 shows an example ofa Profile Dictionary item with its extracted feature – values. Table 5 contains some intermediate results.

Table 5: Example: Intermediate results

sid	25-1	okw	يتضمن
Os:	ويستمر المهرجان لمده ثلاثه ايام بمقر مركز الاطفال المعوقين بحي الملك فهد حيث يتضمن العديد من الانشطه والفعاليات التي تخص الاسره والطفل ومن بينها ندوات تثقيفيه حول اسباب الاعاقه وطرق تلافيها وكيفيه التعامل مع المعوقين		
sPos	lggiggggigghAlifggolggfggggggggggfi		

3.1.2 Creation of relevant Z/D text

Each of the datasets were divided into a set of sentences to be further used in creation, training and testing respectively.

3.2 Training and Evaluation Phase

A similar process to whatwasdone for thedictionary dataset wasalso performed for the examples dataset to prepare it for testing phase. In this phase, the created dictionary was used to extract features for the Example dataset.

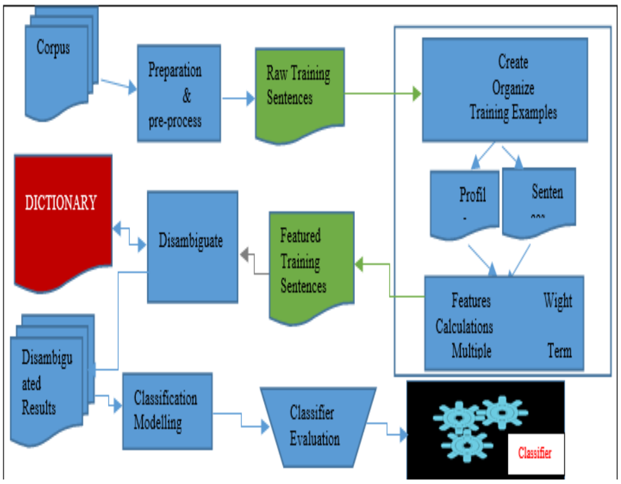

Fig. 2: General Procedure Classification Step 2

3.2.1Feature Calculation and Extraction

In this intermediate phase, the set of procedures dedicated for the calculation and extraction of contextual and other cues wereactivated on the Train and Evaluate dataset. Table 6 contains a sample data record.

3.2.2 Training and Evaluation

In this phase, machine learning classifiers were selected to be trained and tested on the results of the previous phase for each of the sample terms. The following set of classifiers were trained and tested with results as shown in the next section Table 7 and 8.

A cross validation was performed for each sample word using the set of classifiers adopted. A number of important measures were collected for the training and for the validation, including:

- Mean Training Accuracy
- Mean Training Precision, Recall and F1 score
- Mean Training Matthew’s correlation coefficient

- Mean Training Accuracy
- Mean Training Precision, Recall and F1 score
- Mean Training Matthew's correlation coefficientelements in the dataset.

Table 6: Demo of intermediate data on a word pair.

Key	Meaning	THA Word	DHA Word
ID	unique sequential number	677	503
fc	Frequency count	4	1
tty	Type of Word (Z or D)	Z	D
rkw	Root of Word	غيظ	غيض
bkw	Pre Key Context Words	الرغبه في اشفاء المتهم انه اثار العظمي تجذب الحسد	والخسه والنذاله الا
akw	Post Key Context Words	والثار لنفسهم مما وجود الناجح بينهم وقرر قتله والعوده	من فيض وما
okw	Key Word (Original)	والغيظ	غيضا
ows	Key words from examples	ظيغلاو ہظيغ مهظيغي مهظيغ	اضيغ
rwts	TF-IDF Weights for the context roots of key words	اشفاء-0.0 بين-0.0 ثار-0.0 جذب-0.57735 جود-0.0 حسد-0.57735 رغبه-0.0 عظم-0.57735 عوده-0.0 في-0.0 قتل-0.0 قرر-0.0 متهم-0.0 مما-0.0 ناجح-0.0 نفس-0.0 نه-0.0	ؤكد-0.0 اعتبار-0.0 انتخاب-0.40825 اي-0.40825 بيروت-0.0 تحضير-0.40825 ثقه-0.0 حدث-0.0 حكومه-0.40825 خصيص-0.0 خلال-0.0 سن-0.0 طن-0.0 عالم-0.0 مام-0.40825 منح-0.0 وم-0.40825
owts	TF-IDF Weights for the context key words	اثار-0.0 اشفاء-0.0 الحسد-0.57735 الرغبه-0.0 العظمي-0.57735 المتهم-0.0 الناجح-0.0 انه-0.0 بينهم-0.0 تجذب-0.57735 في-0.0 قتله-0.0 لنفسهم-0.0 مما-0.0 والثار-0.0 والعوده-0.0 وجود-0.0 وقرر-0.0	الثقه-0.0 الحكومه-0.40825 امام-0.40825 اي-0.40825 ببيروت-0.0 تخصيصا-0.0 حدثا-0.0 خلال-0.0 طن-0.0 لاعتباره-0.0 لسنتي-0.0 للانتخابات-0.40825 للتحضير-0.40825 منحها-0.0 وعالميه-0.0 يؤكد-0.0 يوما-0.40825
owp	POS of the original words	noun verb noun noun	noun
nsp	Single char signature	lggg gKlglglgfglgalgggiKl laglggNMgglgg	

4. RESULTS AND DISCUSSIONS

To avoid over fitting the testing and evaluation measure were represented by four metrics, TrainingAccuracy with itsMCC values and Validation Accuracy with its MCC using cross-validation. A separate training set wasused to train the model, while another set wasused to do validation to evaluate the model's performance. Training set Metrics are an indication of how the model is progressing in terms of its training. Validation metrics on the validation set measure the quality of the model in its ability to do new predictions based on unseen before data.

Mathews correlation coefficient (MCC) was calculated and used along with overall Accuracy of prediction to analyze the results and select the best classifier. In thiswork a collection from [2] was used and divided into the following data collections. For better and more reliable measurements, the Matthew's correlation coefficient (MCC) wasused. It was considereda more reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset. The mean value across all samples for each measure as well as the max MCC are shown in table 9.

Table 9 and Figures 3 to 6 show the performance of the different classifier models on training and on validation of the sample set of words as well as the MCC for each classifier. Included in the results are the minimum and maximum and the average values obtained by different classifiers.

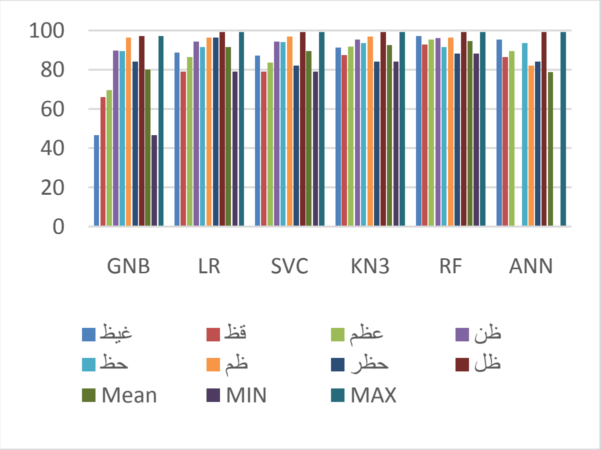

Fig. 3: Sample Words/Classifiers Mean Training Accuracy.

Interms of training accuracies, a number of classifiers performed higher than 90% on two metrics of maximum and average accuracies. Random force, however wasthe champion with highest average and minimum. The maximum was almost the same as the highest value obtained by KNN (94.21 vs. 94.23).This performance of RF was also supported by the highest MCC values as well. The MCC average value and maximum were 0.49 and 0.93 respectively.

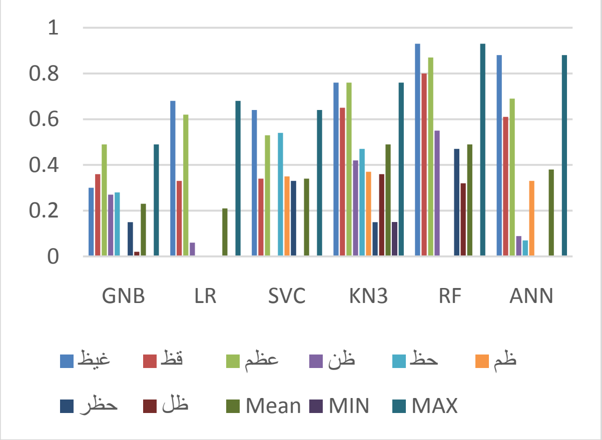

Fig. 4: Sample Words-Classifiers Mean Training MCC

As can be seen from the results table, classifier accuracies varied, but overall, all were reasonably accurate and close in values. MCC values werealso variable with some good ranges. It was notable that some classifiers did not converge and the MCC wasset to zero.

Looking at table 7 which lists a summary of the classifiers with highest training rate and those with highest MCC, it can beeasily conclude that FR was the champion with highest accuracy in most of the terms and many times very close to the highest.

If we consider the mean values across all terms, it becomes obvious that Random Frost wasthe best rate of accuracy and MCC for both training and evaluation.When it comes to validation accuracy and MCC which are more realistic indicators of performance RF was the champion on all three matrices. which means, minimum and maximum were the highest, with values of 99.57%, 84.55% and 99.15 respectively

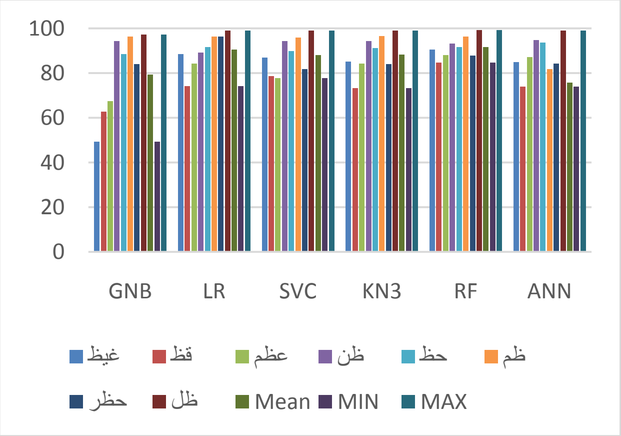

Fig. 5: Sample Words -Classifiers Validation Accuracy

It was also very well supported with highest MCC values of .36 and .74 for the mean and the max respectively.

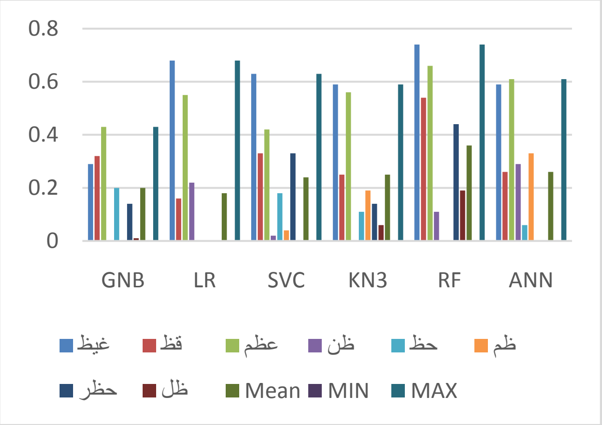

Fig. 6: Sample Words-Classifiers Validation MCC

Table 7: Summary showing RF with best mean Accuracy and MCC

Terms	غيظ		قظ		عظم		ظن		حظ		ظم		حظر		ظل	
Rates	Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation
Accuracy	RF	RF	RF	RF	RF	RF	RF	RF	SVC	SVC	SVC	SVC	LR	LR	KN3	KN3
	97.25	90.58	92.8	84.5	95.4	87.98	96.07	94.71	96.99	89.66	96.79	95.69	96.38	96.38	99.23	99.15
MCC	0.93	0.74	0.8	0.54	0.87	0.66	0.55	0.29	0.8	0.54	0.37	0.19	0.47	0.44	0.36	0.06

Table 8: Mean values across all terms

Mean Training Accuracy	94.60	Mean Training MCC	0.49
Mean Validation Accuracy	91.57	Mean Validation MCC	0.36

Table 9: Training and Validation results of sample words

Measure	**Model**	غيظ	قظ	عظم	ظن	حظ	ظم	حظر	ظل	**Mean**	**MIN**	**MAX**
Training Accuracy	**GNB**	46.7	66.13	69.44	89.77	89.47	96.33	84.08	97.12	**79.88**	**46.7**	**97.12**
	LR	88.77	78.97	86.46	94.28	91.55	96.38	96.38	99.12	**91.49**	**78.97**	**99.12**
	SVC	87.23	78.94	83.54	94.26	93.99	96.79	81.93	99.12	**89.48**	**78.94**	**99.12**
	KN3	91.28	87.53	91.69	95.27	93.64	96.89	84.08	99.23	**92.45**	**84.08**	**99.23**
	RF	97.25	92.8	95.4	96.07	91.55	96.38	88.16	99.21	**94.6**	**88.16**	**99.21**
	ANN	95.45	86.32	89.42	0	93.65	81.93	84.2	99.12	**78.76**	**0**	**99.12**
Training MCC	**GNB**	0.3	0.36	0.49	0.27	0.28	0	0.15	0.02	**0.23**	**0**	**0.49**
	LR	0.68	0.33	0.62	0.06	0	0	0	0	**0.21**	**0**	**0.68**
	SVC	0.64	0.34	0.53	0	0.54	0.35	0.33	0	**0.34**	**0**	**0.64**
	KN3	0.76	0.65	0.76	0.42	0.47	0.37	0.15	0.36	**0.49**	**0.15**	**0.76**
	RF	0.93	0.8	0.87	0.55	0	0	0.47	0.32	**0.49**	**0**	**0.93**
	ANN	0.88	0.61	0.69	0.0889	0.07	0.33	0	0	**0.38**	**0**	**0.88**
Validation Accuracy	**GNB**	49.32	62.57	67.49	94.26	88.48	96.32	84.01	97.06	**79.29**	**49.32**	**97.06**
	LR	88.48	74.19	84.18	89.08	91.55	96.38	96.38	99.12	**90.56**	**74.19**	**99.12**
	SVC	86.81	78.48	77.79	94.22	89.86	95.89	81.73	99.12	**87.99**	**77.79**	**99.12**
	KN3	85.13	73.3	84.21	94.26	91.21	96.5	84.01	99.01	**88.33**	**73.3**	**99.01**
	RF	90.58	84.5	87.98	93.23	91.55	96.38	87.7	99.15	**91.57**	**84.5**	**99.15**
	ANN	84.82	73.93	87.01	94.71	93.64	81.73	84.2	99.12	**75.56**	**73.93**	**99.12**
Validation MCC	**GNB**	0.29	0.32	0.43	0	0.2	0	0.14	0.01	**0.2**	**0**	**0.43**
	LR	0.68	0.16	0.55	0.22	0	0	0	0	**0.18**	**0**	**0.68**
	SVC	0.63	0.33	0.42	0.02	0.18	0.04	0.33	0	**0.24**	**0**	**0.63**
	KN3	0.59	0.25	0.56	0	0.11	0.19	0.14	0.06	**0.25**	**0**	**0.59**
	RF	0.74	0.54	0.66	0.11	0	0	0.44	0.19	**0.36**	**0**	**0.74**
	ANN	0.59	0.26	0.61	0.29	0.06	0.33	0	0	**0.26**	**0**	**0.61**

5. CONCLUSIONS

The problem involving words that contain letters such as (among many others) ظ THA and ض DHA involves what we call near minimal pairs, near homographs (homophones). It requires the determination of the right term and resolutions of created ambiguities. A very amalgamated determination process was suggested that was comprised of multiple stages of feature selection, classifier selection and classification. A sample set of terms wasselected with reasonable success rates making classifier accuracies vary, but overall, all were reasonably accurate and close in values. MCC werealso variable with some good ranges. It wasnotable that some classifiers did not converge and the MCCwas set to zero. Considering results obtained fromclassifiers with highest training rate and those with highest MCC, It can be easily concluded that Random Forest Algorithmwas the champion classifierwith highest accuracy in most of the terms and many times very close to the highest rates, Other classifiers were close. It also scored the highest for the mean value calculation across all terms. It is now clearly seen that with a combination of Extracted features from a corpse along with machine learning classification techniques, the problem can be solved with high accuracy.

6. REFERENCES

[1] Inc., Thinkmap. “Homonym vs. Homophone vs. Homograph on Vocabulary.Com.” Homonym vs. Homophone vs. Homograph: Choose Your Words | cabulary.Com, www.vocabulary.com,https://www.vocabulary.com/articles/chooseyourwords/homonym-homophone-homograph/. Accessed 6 Aug. 2022.

[2] Abu El-khair, I. (2016). Abu El-Khair Corpus: A Modern Standard Arabic Corpus. International Journal of Recent Trends in Engineering & Research, 2(11), 5-13,

[3] Ide, N., Véronis, J., (1998) “Word Sense Disambiguation: The State of the Art”, Computational Linguistics, Vol. 24, No. 1, Pp. 1-40.

[4] Cucerzan, R.S., C. Schafer, and D. Yarowsky, (2002) “Combining classifiers for word sense disambiguation”, Natural Language Engineering, Vol. 8, No. 4, Cambridge University Press, Pp. 327-341.

[5] Nameh, M. S., Fakhrahmad, M., Jahromi, M.Z., (2011) “A New Approach to Word Sense Disambiguation Based on Context Similarity”, Proceedings of the World Congress on Engineering, Vol. I.

[6] Xiaojie, W., Matsumoto, Y., (2003) “Chinese word sense disambiguation by combining pseudo training data”,Proceedings of The International Conference on Natural Language Processing and Knowledge Engineering, Pp. 138-143.

[7] Navigli, R. (2009) “Word Sense Disambiguation: a Survey”, ACM Computing Surveys, Vol. 41, No.2, ACM Press, Pp. 1-69.

[8] “Minimal Pairs Theory.” Minimal Pairs Theory, www.speechlanguage-resources.com, http://www.speechlanguage-resources.com/minimal-pairs-theory.html. Accessed 5 Aug. 2022.

[9] Barlow, J.A. and Gierut J.A. (2002) Minimal Pair Approaches to Phonological Remediation Seminars in Speech and Language, Volume 23, No 1

[10] Bowen, C. (2009) Children's Speech Sound Disorders Wiley-Blackwell

[11] Williams, A.L. McLeod, S. & McCauley R.J. (2010) Interventions for Speech Sound Disorders in Children Paul H Brookes Publishing Co

[12] Williams, A.L. (2006) SCIP Sound Contrasts in Phonology: Evidence Based Treatment Program. User Manual Super Duper Publications

[13] Carpaut, M., and Wu, D., 2005, Word Sense Disambiguation vs. Statistical Machine Translation”, in Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 387-394.

[14] Chan, Y., Ng, H., and Chiang, D.,2007, “Word Sense Disambiguation Improves Statistical Machine Translation”, in Proc. of the 45rd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 33-40.

[15] Schütze, H., and Pedersen,1995, “Information Retrieval Based on Word Senses”, in Proc. of Symposium on Document Analysis and Information Retrieval (SDAIR’95), pp. 161-175.

[16] Stokoe, C., Oakes, M., and Tait, 2003, “Word Sense Disambiguation in Information Retrieval Revisited”, in Proc. of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 159-166.

[17] Atkins, Sue. 1991. Tools for computer-aided corpus lexicography: The Hector project. Acta Linguistica Hungarica, 41: 5–72.

[18] Jacquemin, B., Brun, C., and Boux, C.,2002, “Enriching a Text by Semantic Disambiguation for Information Extraction”, in Proc. of the Workshop on Using Semantics for Information Retrieval and Filtering in the 3rd International Conference in Language Resources and Evaluation (LREC).

[19] MALLERY, J. C. 1988. Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers. Ph.D. dissertation. MIT Political Science Department, Cambridge, MA.

[20] NG, T. H. 1997. Getting serious about word sense disambiguation. In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? (Washington D.C.). 1–7.

[21] Elhadi, M. “Arabic News Articles Classification Using Vectorized-Cosine Based on Seed Documents.” (2019).

[22] Edmonds P, Agirre E. Word Sense Disambiguation: Algorithms and Applications.

[23] Manzour, I. (2017) Lisan Al-Arab. www.lesanarab.com

[24] El-Gamml MM, Fakhr MW, Rashwan MA, Al-Said AB. A comparative study for Arabic word sense disambiguation using document preprocessing and machine learning techniques. InArabic Language Technology International Conference, Bibliotheca Alexandrina, CBA (Vol. 11).

[25] Zouaghi A, Merhbene L, Zrigui M. A hybrid approach for arabic word sense disambiguation. International Journal of Computer Processing Of Languages. 2012 Jun;24(02):133-51.

[26] Saad MK, Ashour WM. Osac: Open source arabic corpora. In6th ArchEng Int. Symposiums, EEECS 2010 (Vol. 10).

[27] Diab M, Alkhalifa M, ElKateb S, Fellbaum C, Mansouri A, Palmer M. Semeval-2007 task 18: Arabic semantic labeling. InProceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007) 2007 Jun (pp. 93-98).

[28] Dixit V, Dutta K, Singh P. Word sense disambiguation and its approaches. CPUH-Research Journal. 2015;1(2):54-8.

[29] Albogamy F, Ramsay A, Ahmed H. Arabic tweets treebanking and parsing: A bootstrapping approach. InProceedings of the Third Arabic Natural Language Processing Workshop 2017 Apr (pp. 94-99).

[30] Elmougy S, Taher H, Noaman H. Naïve Bayes classifier for Arabic word sense disambiguation. Inproceeding of the 6th International Conference on Informatics and Systems 2008 Mar 27 (pp. 16-21).

[31] Chalabi A. Sakhr Arabic-English computer-aided translation system. InConference of the Association for Machine Translation in the Americas 1998 Oct 28 (pp. 518-521). Springer, Berlin, Heidelberg.

[32] Hakak SI, Kamsin A, Shivakumara P, Gilkar GA, Khan WZ, Imran M. Exact string matching algorithms: Survey, issues, and future research directions. IEEE access. 2019 Apr 30;7:69614-37.

[33] Merhbene L, Zouaghi A, Zrigui M. Lexical Disambiguation of Arabic Language: An Experimental Study. Polibits. 2012 Dec (46):49-54.

[34] Diab M, Resnik P. An unsupervised method for word sense tagging using parallel corpora. InProceedings of the 40th Annual Meeting of the Association for Computational Linguistics 2002 Jul (pp. 255-262).

[35] Eid MS, Al-Said AB, Wanas NM, Rashwan MA, Hegazy NH. Comparative study of rocchio classifier applied to supervised wsd using arabic lexical samples. InProceedings of the tenth conference of language engeneering (SEOLEC’2010), Cairo, Egypt 2010 Dec 15.

[36] Mirtaheri SL, Shahbazian R. Machine Learning: Theory to Applications. CRC Press; 2022 Sep 29.

[37] Biau G, Scornet E. A random forest guided tour. Test. 2016 Jun;25(2):197-227.

[38] Brereton RG, Lloyd GR. Support vector machines for classification and regression. Analyst. 2010;135(2):230-67.

[39] Kaur G, Oberai EN. A review article on Naive Bayes

classifier with various smoothing techniques. International Journal of Computer Science and Mobile Computing. 2014 Oct;3(10):864-8.

[40] Taunk K, De S, Verma S, Swetapadma A. A brief review of nearest neighbor algorithm for learning and classification. In2019 International Conference on Intelligent Computing and Control Systems (ICCS) 2019 May 15 (pp. 1255-1260). IEEE.

[41] Wang QQ, Yu SC, Qi X, Hu YH, Zheng WJ, Shi JX, Yao HY. Overview of logistic regression model analysis and application. Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine]. 2019 Sep 1;53(9):955-60.

[42] Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H. State-of-the-art in artificial neural network applications: A survey. Heliyon. 2018 Nov 1;4(11): e00938.

[43] Zerrouki, T. Tashaphyne, Arabic light stemmer, 2012. https://pypi.python.org/pypi/Tashaphyne/0.2

[44] Spärck Jones, K. (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". Journal of Documentation. 28 (1): 11–21. CiteSeerX 10.1.1.115.8343. doi:10.1108/eb026526.

[45] Obeid O, Zalmout N, Khalifa S, Taji D, Oudah M, Alhafni B, Inoue G, Eryani F, Erdmann A, Habash N. CAMeL tools: An open source python toolkit for Arabic natural language processing. InProceedings of the 12th language resources and evaluation conference 2020 May (pp. 7022-7032).https://pypi.org/project/strsimpy/