Translating Ambiguous Arabic Words using Text Mining

Omer Awad Mohammed
Department of Computer Science
Northern Border University
Saudi Arabia

Ahmed Salah
Professor, Dean of Faculty of Computer Science
Sudan Open University
Sudan, Khartoum

ABSTRACT
Language is a fundamental value in the life of every nation, it is a tool that carries ideas and concepts. Languages are a means of communication between humans is through word means clear .Languages of various kinds have the controls and rules need to be understood we can deal with them.

Arabic language is considered a pioneer of languages and it's enough recognition language of the 'Koran' and in a addition to that spoken by millions around the world.

Arabic is the Language of the vast and complex language, Arabic script can be has several exists, because every Arab character has many express movements and this is what distinguishes the Arabic languages from other languages.

General Terms
Ambiguous, Text mining

Keywords
Ambiguous Words, Data Mining , Text Mining

1. INTRODUCTION
Arabic language is considered to be one of the languages with wide scope and different branches ranging from Rhetoric, Grammar, and syntax.

There are many words in Arabic language does not translate properly, such as the **(وسام)**, this name but he has more than one meaning may be the Order of Courage, or the name of a person, so you need to be identified and one meaning true.

In general words that have meaning adjective and at the same time be the name of a person. We need to identify one meaning to this kind of words.

This is of the challenges facing the Arab language now, the ambiguous words which have more than one meaning. Researcher aims to take advantage of data mining technology to meet this challenge.

Text mining (TM), known as textual data mining techniques, refers to the processes of extracting or discovering knowledge from a text. The goal of TM is to discover unknown information which can help an individual make better decision in the future based on textual data.

2. METHODOLOGY
This paper aims to describe a methodology to identify ambiguous words tendencies in a classification process. The main purpose of text mining is dialed with unstructured data which data write in natural language and have more ambiguity. Researcher working on the development of a methodology to solve the problem of ambiguous words and so through the following:

a.Tokenization *:* The first process in text preprocessing is transforming the text from a stream of words by replacing all non-text letters such as numbers, punctuation marks (!),(?),(,),(:),(.), (،)…etc. and any other characters such as @,#,$,%,^,&,*,>,<….etc. to single spaces. Also, these tokenization technique results are used for further processing.

B.Filtering*:* In this step, we filter all of the characters not related to the Arabic language and the Arabic stop words that are infrequent and un-useful in text classification. Also, some Arabic words that are less than three letters are removed , then represent the textual data in matrix form to illustrate the weight of each term or feature in each text.

C.Text Pre-processing*:* Text preprocessing is a basic step before text classification. It is applied to prepare the text by transforming it into a suitable format to be able to apply different textual data mining techniques.

D.Classification: Researcher suggests the classifiers that help to understand the "general tendency" of a word/group of words in training text and have better performance. Then, they will be used to calculate the "particular tendency" of word/group of words for a given text in the testing phase. The classification part has two phases: training and testing. Phase training is used in this experiment by entering the labeled text and learned models to find the best performance model and to analyze the ambiguous words. The testing phase is used to apply the model that gains the best performance and to test it on a new text.

All these models will be applied in the training phase first to learn the classification and to assess the performance of the classifier algorithms results. Therefore, we will choose the best classifier that gives the highest accuracy of classification to apply it in classification and to find the particular tendencies of word. Researcher seeks to be a tendency process is a neural network is tested to get the best results.The researcher used the method Rider based on Edit Distance (RED) , this method automatically computes the score related to the translated output of the machine translation system using a decision tree (DT).

3. PROBLEM STATEMENTS
The paper discusses the following problems:

A.The concept of ambiguous words in Arabic language by clarifying the concept of the words that have more than one meaning (a word meaning name and another meaning adjective) or the meaning of the name and the another meaning of an verb and to prepare a list containing the largest number of ambiguous words in the Arabic language

B.Processing ambiguous words and translate them to choose one meaning correctly using the Data Mining Technology.

4. OBJECTIVES
This paper aims to follows:

1- Understand the meaning of ambiguous words in the Arabic language.

2- Enhance the effectiveness translator to get the best results (translation ambiguous words).

3- Find a standard approaches to understand the general direction of the word (general tendencies).
4- Built Dictionary that contains a large number from the list of ambiguous words in the Arabic language and clarify the correct translation of these words.
5- Implementation of the three previous points by designing a system to clarify the mechanism of correct translation.

5. LITERATURE REVIEW

5.1. The Title Of The First Study Is: Identifying Word(S) Tendencies In Classification Of Arabic News Using Text Mining Techniques.

Amani, Al-Ghanayem, Dr.Waleed, Rashaideh College of Computer and Information Science, Al-Imam Muhammad Ibn Saud Islamic University, Saudi Arabia College of Computer and Information Science, Al-Imam Muhammad Ibn Saud Islamic University; Saudi Arabia.

5.1.1 Objective of the study

Describe the methodology for determining the ambiguous words tendencies.This study produces a methodology using classification models to identify the word(s) tendencies in Arabic news text towards certain categories.The main challenge for analyzing and classification Arabic texts is that they have more ambiguous word (the same word has different concepts in different contexts) for example the word (ذهب) can refer to different concepts but is spelled the same as a noun , it means gold and as a verb , it means go .

5.1.2 The study methodology

This study defined a methodology into two parts:

5.1.3 The classification part has two phases

Training and testing. Phase training is used in this experiment by entering the labeled text and learned models to find the best performance model and to analyze the ambiguous words.

5.1.4 Association mining part

Seek to find the relationships between words and their tendency to towards different categories and to eliminate ambiguous words**.**

This study has prepared Dictionary specific number of ambiguous words in Arabic and translated. Researcher works to add a large number of ambiguous words in Arabic and translated.

The study defines estimates the general direction of the word (general tendencies), but the researcher is working to illustrate a standard approach for the general direction of the word.

5.2 The Second Study

Evaluating English to Arabic Machine Translation Using BLEU.(IJACSA) International Journal of Advanced Computer Science and Application , Vol.4,No.1,2013.One of the methods used to evaluate machine translation system is Bilingual Evaluation Understudy (BLEU) which was introduced in the study of Papineni , Roukos, Ward and Zhu [1].

5.2.1. Objective of the study

To compare the effectiveness of two popular machine translation systems (Google Translate) and Babylon machine translation system [2]) .

5.2.2. The study methodology

The methodology followed the main steps, in the first step input five statements; the source sentence in English is inputted to machine translation system. The translation of the source sentence using Google Translate System. The translation of the source sentence using Babylon Translate System . Two reference translation of the source sentence. The second step : Handling text by dividing it into different weights are (n-gram)n-gram is a sub-sequence of n elements of a particular sequence of words.The following figure illustrates N-grams Extraction Flowchart.

5.2.3 BLEU problem

lacks some of the characteristics, if the word more than one meaning; for example (Shaker شاكر , Omer عمر) . Therefore a number of researchers have attempted to enhance this study. One of such attempts was conducted by Yang et al. [3] , use (Multiple Linear Regressions) to assign proper weights to different n-grams and words within BLEU framework. Enhance the effectiveness of the translator here researcher working on the use of Rider based on Edit Distance (RED) , this method automatically computes the score related to the translated output of the machine translation system using a decision tree (DT).

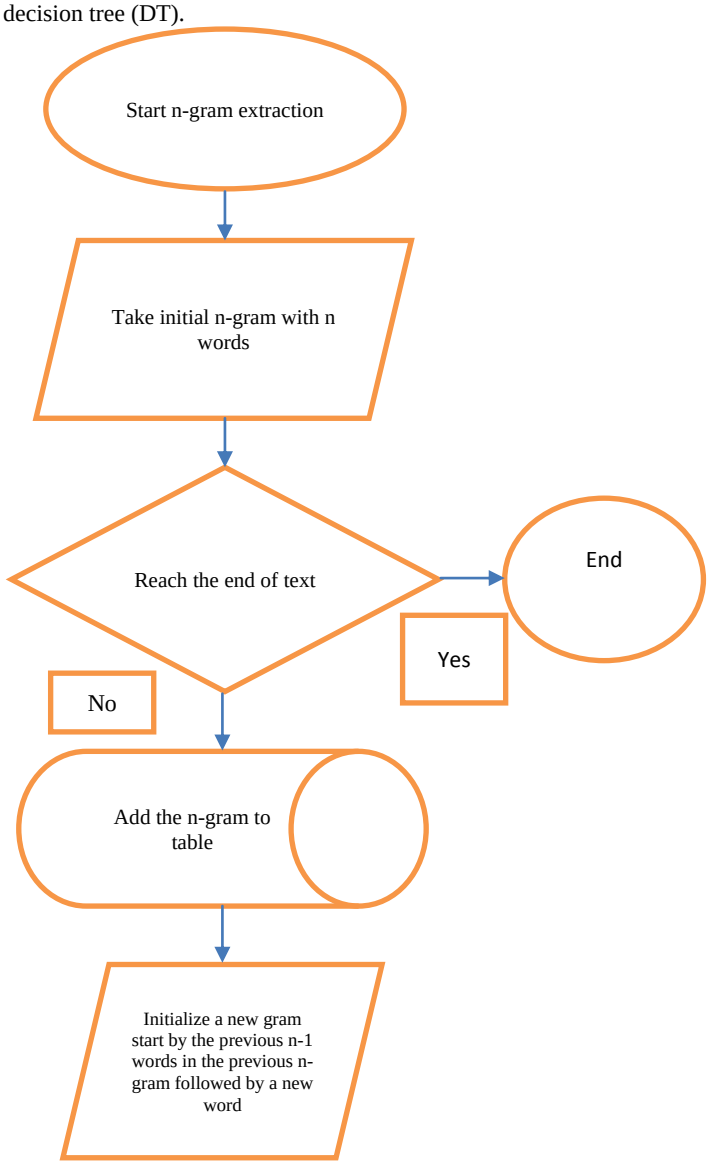

Figure: N-grams Extraction Flowchart

5.3 The Third Study: Google Translate

Translate words here in several languages such as translation from Arabic to English or vice versa. Google Translate does not apply grammatical rules, since its algorithm are based on statistical analysis rather than traditional rule based analysis.

5.3.1 The study methodology

Google translate based on Statistical Machine Translation (SMT) , is a machine translation paradigm where translations are generated on the basis of statistical models .

5.3.2 The main problem with Google

Translate are Inadequacy of the recorded corpus of language that is used for reference , the corpus is limited to past forms of the language. The quality of Google translate and machine translate in general depends on the similarity of the languages . If there is more than one meaning of the word cannot find specific translation , for example (وسام) Translated (Decoration), (جمال) Translated (Beauty).Researcher working on the preparation of a system to deal with ambiguous translation of certain words

5.4 The Fourth study Address : Using Fuzzifiers to Solve word Sense Ambiguation in Arabic Language

This study was presented by : Madeeh Nayer EL-Gedawy (Computer Center ..Institute of Public Administration (IPA) – Jeddah . Published (International Journal of Computer Applications (0975-8887)October 2013

Text mining techniques confront many challenges when dealing with the Arabic language including lexical disambiguation because Arabic is a highly inflectional and derivation language. This study is treating Word Sense Disambiguation (WSD) as a pure text classification problem by marking the correct sense using keyword in context. (WSD) is one of the trickiest tasks in text mining.

5.4.1 Objectives of the study

The main objectives of this study are:

The fuzzy logic membership function that is used in allocating words to senses .

Creating an Arabic sense inventory out of the English WordNet instead of depending on Arab WordNet which has very poor coverage.

Enriching the training set derived from the knowledge base by extending the sense inventory through query expansion.

5.4.2 Methodology of the study

This study followed a methodology: to fulfill the task of word sense disambiguation, firstly begin by some preprocessing tasks in Arabic:

They are two important processing tasks in this section

First : stop word removal

a stop word is defined using two criteria : first , it must have a high frequency of document document(DF). Second , the terms correlations with categories should be small. Word frequency is the number of times a word appears in a document .

Second : Root extraction and Word Stemming

There are two types of stemmers: root extractors and light stemmers. Root extractors are aggressive stemmers that confront the problem of over-stemming where many words of different meanings can be conflated to the same root. For example, in the verse: " فسينفقونها ثم تكون عليهم حسرة ثم يغلبون"; we notice that the word 'فسينفقونها' if expressed in a 3 letters root 'نفق', then we will have 4 different meanings: ' نفق لعبور السيارات','نفق أى مات','أنفق المال', and ' نفاق ومنافق'. So, over-stemming leads to many candidates that should be examined carefully and that leads to a more complex analysis. On the other hand, light stemmers try to find the shortest possible path without compromising the meaning, so it limits the candidates as much as possible but it sometimes fails to deal with affixes and broken (irregular) plurals.

5.5 Fifth Study:Address: The Role of Ambiguity in Arabic Language

Presented by Hamza Al- Harbi **.** Department of English University Sains Malaysia

Published **:**The International Journal of Social Sciences and Humanities Invention Volume 3 issue 6 2016 page no.2222-2227 ISSN: 2349-2031

This study highlight the vital role of ambiguity in language. Ambiguity led to Misunderstanding that cause a breaking down of the relationship among communicators. To avoid these mistakes that occur as result of ambiguity the learners should to know how to solve these issues. This study explains how to solve these problems by providing some examples from Arabic language. To sum up, the ambiguity in Arabic language is more problematic than others. By providing some example on how disambiguate these sentences the learners of Arabic as foreign language will achieve the goal of communication.

5.5.1 Objectives

The aim of this study is to solve ambiguity in some Arabic words by clarifying some ambiguous words in the Arabic language and then clarifying the mechanism of treatment.

5.5.2 Method of Disambiguation in Arabic language

Many of the ambiguities can be resolved by looking at the context. The linguistic contexture can resolve many of the ambiguities especially among different word classes From the development point of view, processing and disambiguation of Arabic depend in the following sources of information:

a- The lexicon: provides basic and initial information about lexical items (grammatical attribute).
b- Adjacency constraints: specify the compatibility or the incompatibility of two neighboring.
c- The Idafa (The IDAFA construction is an important grammatical structure in Arabic. It is a genitive construction in which two nouns are linked in such a way that the second (second part of the construction) qualifies or specializes the first (first part of the construction) construct cannot be followed by a preposition.

5.5.3 An example of ambiguity resolution

ذهب they went (verb) or gold (accusative) .The disambiguation process is started by using the adjacency condition that a noun cannot be followed by a preposition الى to. Thus, (ذهبا they went) is a verb (go) [MASC, DUAL} not a noun. Sami (سامي) is a named entity cannot be the subject of the verb as there are no morphological dependencies (agreement in number). On

6. CONCLUSION

This research aims to solve the problem facing the Arab language, ambiguous words, Researcher work to take advantage of data mining technology to solve this problem. Researcher deal with a list containing a large number of ambiguous words in the Arabic language and the work methodology used Data Mining technique to determine the one meaning .

7. REFERENCES

[1] Dr. Mohammed N. Al-Kabi Faculty of Sciences & IT Zarqa University Zarqa Dr. JordanTaghreed M. Hailat, Emad M. Al-Shawakfa, and Izzat M. Alsmadi Faculty of IT and CS Yarmouk University. (2013) . Evaluating English to Arabic Machine Translation Using BLEU

[2] Bonnie Glover Stalls and Kevin Knight. USC Information Science Institute bgsoisi.edu,knightoisi.edu (2013 Translating Names and Technical Terms in Arabic Text.

[3] Dr. Barihi Adetunji . National , Dr A.Raheem Mustapha Open University of Nigeria School of Science (2013). Translation (Arabic – English).

[4] Argamon, Krymolowski . International Conference on Computational Linguistics. A memory – based approach to learning shallow Natural Language Patterns .

[5] K.R. Chowdharry , Professor and head CSE Deprt.MBM. Engineering colledge , Jodhpur,India (2013). Natural Language Processing.

[6] Dr.Vishal Gupta Computer Science & Engineering, University Institute of Engineering & Technology, Panjab University Chandigarh, India. Gurpreet S. Lehal Professor & Head, Department of Computer Science, Punjabi University Patiala, India(2014). Text Mining Techniques and Applications.

[7] Dr. Nazik Abdel-Lateefph.d wales University , UK and Dr. Iman Adawyph.dBanha University (2012/2013). Translation from and into Arabic

[8] Matthias Huck and David Vilar . Human Language Technology and Pattern Recognition Group, RWTHAachen University. Advancement in Arabic to English Hierarchical Machine Translation.

[9] Dr.AbdelKarim Mohammed . An-Najah National University , Nablus , Palestine, March (2014). Translating contracts between English and Arabic.

[10] A. Cheung, M. Bennamoun*, N.W. Bergmann Space Centre for Satellite Navigation, School of Electrical & Electronic Systems Engineering, Queensland University of Technology (2014). An Arabic optical character recognition system using recognition-based segmentation

[11] Clare Brierley, School of Computing, University of Leeds Majdi Sawalha, Computer Information Systems, University of Jordan,Barry Heselwood, Linguistics and Phonetics.