Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.199

IJCSMC, Vol. 8, Issue. 2, February 2019, pg.130 – 140

Translating Ambiguous Arabic Words Using Text Mining

Omer Awad Mohammed
Graduate College of Computer Science, Alneelian University, Khartoum, Sudan
umerawad@hotmail.com

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

ABSTRACT: Ambiguous words in the language translate into more than one meaning, Ambiguity is a problem that to be processed. The researcher reviews many of studies that discussed the ambiguity of words in language with the results.
The researcher aims to find a solution to the problem of ambiguity using artificial intelligence techniques.
General Terms: Ambiguous words, Artificial Intelligence Techniques

INTRODUCTION

Arabic language is considered to be one of the languages with wide scope and different branches ranging from Rhetoric, Grammar, and syntax. 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 researcher presents twelve studies discussing ambiguity in language with the definition of a solution to this problem.

1- 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).

Objective of the study is:

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.

The study methodology:

This study defined a methodology into two parts:

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.

Association mining part: work 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.

The study defines estimates the general direction of the word (general tendencies).

2-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 Paining, Roukos, Ward and Zhu [1].

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 use (Multiple Linear Regressions) to assign proper weights to different n-grams and words within BLEU framework.

Enhance the effectiveness of the translator, 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).

3- 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.

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.

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).

Objective of the study:

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

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.

4- 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)

Abstract:

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.

Objectives:

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.

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: 'أنفق المال','نفق أى مات','نفق لعبور السيارات', 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- The Role of Ambiguity in Arabic Language

Presented by (Hamzeh 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

Abstract:

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.

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.

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.

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.

An example of ambiguity resolution

ا + 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 the other hand, a morphological dependencies exists between ذهبا and صاحبا suggesting that it is (two friends) and that it is the subject. This solution is verified by the existence of a morphological dependency between صاحبا two friends) and سامي Sami: the suffix that indicates duality ending is (NOM), but when the noun is the first part of the IDAFA construction. The suffix should be which the case is in the above sentence. So, Sami is the second part of the IDAF construction.

6- Machine Translation of Arabic Language: Challenges and Keys

(Eltayeb Abuelyaman. Department of Computer Science (University of Nizwa))

(Wafa Mukhtar, Limia Rahmatallah, Mona Elagabani (Department of Computer Science) (Sudan University of Science & Technology))

(Fifth International Conference on Intelligent Systems, Modelling and Simulation)

The study discussed challenges for researchers in Arabic language, the challenges are very clear when translating from Arabic to English using machine translation today, for two reasons:

- The first is the impenetrability of some Arabic words, which puzzles their decomposition into morphemes.
- The second is the incompatibility of existing machine translation techniques with the Arabic language.

These challenges are related to the most advanced technology available today.

This study proposes a framework for the Arabic translation process, also, it recommends the all-inclusive Holey Quran's English translation as the most reliable benchmark for Arabic to English Translation.

To contribute to the translation from Arabic to English using machine translation, this study proposes two recommendations:

- The first is to develop a two-phase which analyzes an Arabic text during a top-down phase and then translates it into English during the bottom-up phase.
- The second is to adopt a standard benchmark for assessing performance of the first

The Top-Down Phase:

The top-down phase is carried out by a process that analyzes Arabic texts. It basically decomposes a text into its atomic components. Analysis is the only top-down stage.

- **The Bottom-Up Phase** :
  The bottom-up phase includes the following stages:
  a- Mapping
  b- Permutation
  c- Synthesis

The study discussed the concept of compound words, the compound words in Arabic are translated into the following stages:

First: The compound word is divided into sections.

Second: Translation of each section of the compound word into English letters.

Third: Translation from Arabic into English.

The following example explains translation in compound words:

word	و بقولهم			
components	هم	قول	ب	و
transliteration	himm	qouli	bi	wa
translation	their	say	by	and

The study also discussed Subject embedding:

The Arabic language deals with pronouns, since the absent pronouns do not write in the sentence but in the English language is written in the sentence. For example, the statements **"He ate a chicken**" can expressed in Arabic as "**اكل دجاجة".** The subject "**He**" and the verb **"ate**" are represented in Arabic by the single verb-form **"اكل**". That is; **"He ate** "is translated **as (اكل**) and **"a chicken** "is translated as **(دجاجة**).

The study identified some concepts and explained them in examples:

- **Relaxed Sentence Order** : Arabic exhibits a large degree of freedom in the order of words within a sentence .For example ; the sentence (The men ate a bull) ; Can be translated in Arabic to:

(الرجال اكلوا ثورا) or (اكل الرجال ثورا)

- **Word Ambiguity** : the study introduce some examples of words ambiguity like Arabic word (**خال**), which can be translated to any of the following three words "**empty**", **imagined**", "**battalion**"

Note (This study did not suggest a specific mechanism for processing such ambiguous words.

- **Letter ambiguity**: Ambiguity is not limited to Arabic words only. Some Arabic letters when affixed to morphemes lead to ambiguous compound words .for example the letter "ب" takes on any of the following sense :through , in , by , for and at.

The table shows the translation of letter (ب)

word	translation	Word (ب)	Translation of word (ب)
بركة	Blessing	ببركة	Through blessing
البيت	The house	بالبيت	In the house
المال	The money	بالمال	By the money
اي	What	باي	For what
الباب	The door	بالباب	At the door
القلم	The pen	بالقلم	Using the pen

Study Summary:

This study discussed the challenges of translation in Arabic. The study defined compound words, words ambiguity and letter ambiguity.

The study did not specify a specific mechanism for processing ambiguity in words and ambiguities

7- Arabic morpho-syntactic feature disambiguation in a translation context

(Ines Turki Khemakhem, Salma Jamoussi, Abdel Majid Ben Hamadou)

(MIRACL Laboratory, ISIM Sfax, Pole Technologies)

Note:

The study used translation from Arabic to French.

This study focused on the concept Morphological disambiguation in Arabic language, (the morphological analysis of a word consists of determining morphological information about each word).

The study proposes a system that eliminates ambiguity by using morphological, the stages of this system are:

- Segmentation of Arabic words.
- Using the morphological analyst (MOROPH)
- Analysis of the word in the morphological analyst using a tree structure.

Summarize the study:

The basic idea of study is to process ambiguity. The study used an analyst Morpho, Which works to divide the Arabic word into segments, then analyze them. A tree structure is used to analyze and translate the word. The study discussed translation of the Arabic words into French.

8- Arabic English Word Translation Disambiguation using Parallel Corpora and Matching Schemes

(Farag Ahmed and Andreas Nurnberger [Data and Knowledge Engineering Group – Faculty of Computer Science – Otto-von-Guericke-University of Magdeburg])

Objectives:

Describe the implementation and evaluation of an Arabic English word translation disambiguation approach that is based on exploiting a large bilingual corpus and statistical co-occurrence to find the correct sense for the query translations terms.

The study discussed a concept of Word Sense Disambiguation (WSD), the meaning of a word may vary significantly according to the context in which it occurs. As a result, it is possible that some words can have multiple meanings. This problem is even more complicated when those words are translated from one language into others.

Study defined (WDS), in general, is the process of determining the right sense of an ambiguous word given the context in which the ambiguous word occurs.

The study proposed several mechanisms to process ambiguity, such as (Discretization) in Arabic, sometimes called vocalization. It is a symbol placed on the letters in the Arabic word, used to indicate the correct meaning of the word. For example, the Arabic word "**يعد**" can have these translation in English (Promise , Prepare , count , return , bring back) or the Arabic word " **علم**" can have these possible translations (flag , science , he knew , it was known , he taught , he was taught)

Methodology of the study to translate the ambiguity words:

For example the Arabic words (**يعد ، علم)**

First: identifying all senses for every word relevant, by using a list of senses for each of the ambiguous words.

Second: assign each time this word occurs the appropriate sense to it, by analyzing of the context in which the ambiguous word occurs, or by use of an external knowledge source, such as lexical resources.

The study was classified (WSD) according to machine learning into three categories: supervised learning, unsupervised learning, and combinations of them.

The study discussed a concept of Corpora. It is a mechanism that relies on collecting information from the data to clarify the specific meaning of the ambiguous word.

The study described a method based on Naïve Bayesian Algorithm, The word attributes are represented in the form of variables for possible meaning.

Study Summary:

This study discusses the concept of ambiguity in the Arabic word by clarifying the meaning (WSD), Using two examples of words that are ambiguous in Arabic **(يعد ، علم)**.

The study defined the method of disambiguation by studying and analyzing the context in which the word is ambiguous, the text is analyzed using a statistical equation (Naïve Bayesian).

9- Handling Text Mining Problems in Arabic using Domain – Specific Approach

(Madeeh Al-Gedawy, Osman Hehazy; Department of Information Systems, Cairo University, International Journal of Computer Applications (0975-8887))

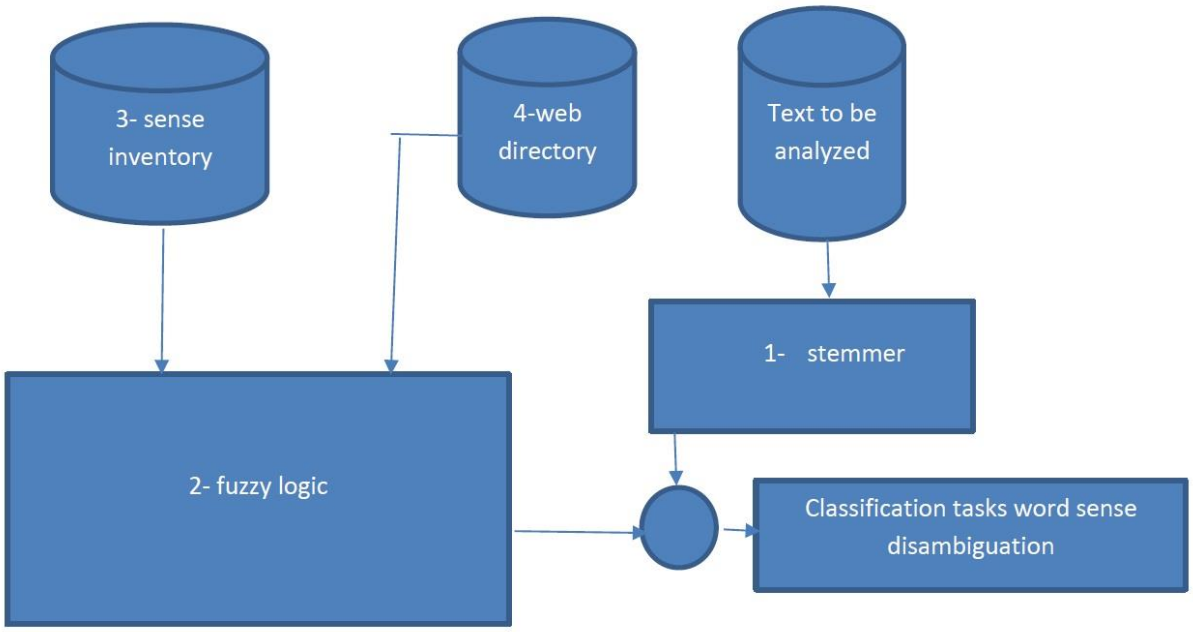
Study objective:

The study aims to solve the problem of ambiguity in Arabic words, describes resolution of ambiguity by using Word Sense Disambiguation. The ambiguous word translates according to the characteristics of the word.

The study methodology

The study uses a special approach called (Domain Specific Class), the Arabic text is analyzed according to the description and form. For example: can say (فتاة رقيقة) but cannot say (سيارة رقيقة).

The study described a system for translating ambiguous words, as in the following diagram:



Fuzzy logic:

Fuzzy logic provides a means for encapsulation the subjective decision making process in algorithm suitable for computer implementation.

Algorithm Description:

Suppose that the context "c" contains the ambiguous word "w" and that the context has a window of "k" words and that w has "n" senses so for fuzzification; fuzzy set is to be defined for every word sense "S" So that the fuzzy set (FS) of sense "y" is

FS (Sy) = U [D (w1), D (w2), D (wk)]

Where D (w1) is a membership for the first word w1 in the context 'c' to be allocated to the sense 'y'; the U is the union operator, in general the union operator in fuzzy logic is equivalent to max operator.

Study Summary:

The study used domain – specific classifier fuzzy logic for translation. The Arabic domain-specific classifier technique enable the researchers to accurately classify and disambiguate Arabic words .The technique proved to better results in classification and word sense disambiguation.

10- Divergence and Ambiguity Control in an English to Arabic Machine Translation

Marwan Akeel et al Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 3, Issue 6, Nov-Dec 2013, pp.1670-1679 (www.ijera.com)

Objective of the Study:

Translate ambiguous words in English language into Arabic language by using machine translation.

This study proposed a strategy for detecting and controlling differences and ambiguity in the translation system, this strategy includes tools and data structures in the language dictionary. The study was used ANN as a bilingual dictionary.

There are eight main modules on our system (translation system):

1- **Sentence Separator and Contractions Removal**: This module first separates the paragraph into sentences. Then each sentence is processed. If any contraction is present in the sentence, it is removed.
2- **Parser and Tagger:** This module divides the sentence into a tree structure and then derives grammatical relations between the sentence words.
3- **Knowledge Extraction**: This module extracts information from parser and tagger for each part of the sentence. Each part of the sentence is converted to knowledgeable object by adding all the information associated with it and sentence is represented as a collection of knowledgeable objects.
4- **Grammar Analysis and Sentence Structure Recognition**: This module processes the collection of knowledgeable objects and recognizes parts of the sentence: subject, main verb, auxiliary verb, object, indirect object. Each recognized part of the sentence contains one or more word called chunk. Each sentence is divided into chunks. The grammatical attributes of the sentence such as tense, voice, type, and form are also identified. This module generates the grammatical structure of the English sentence by analyzing the chunks and attributes of the sentence. It also decides the Arabic sentence structure and word order according to the English recognized sentence structure.
5- **ANN and Rule based mapping**: The ANN module looks for each word in the bilingual dictionary object which is trained for word mapping, and gets the corresponding Arabic word and associated information in numeric form. This result is decoded to textual form by the Encoder-Decoder.
6- **Words Selection and Modification and Syntax Addition:** In this module select the correct form out of the received Arabic translation of each chunk word. The selection is done depending on the sentence tense or the features of the chunk main word or the subject main word.
7- **Arabic Sentence Generation:** In this Module, The meanings are arranged according to the form of the Arabic sentence obtained in module (Words selection and Modification).
8- **ANN Module:** used feed-forward back-propagation artificial neural network for the selection of Arabic words/tokens (such as verb, noun/pronoun) equivalent to English words/tokens. Each English word is matched to at least two meanings in Arabic.

Example:

My friend helps the old man يساعد صديقي الرجل العجوزا (Ysaed Sdygy Alrjl AlEjwz)

يساعد صديقي الرجل القديم(Ysaed Sdygy Alrjl Alqdym)

The word (0ld) translated into two words ((العجوز,) (القديم. The word ((القديم does not suit the man in the Arabic language, in this case, the form must be analyzed to extract (verb and object). The translation system uses a statistical analyzer to derive the most probable sentence analysis of the correct meaning, and finding grammatical relations between words.

Study Summary:

This study discussed ambiguity in translating words from English into Arabic, It proposed a system with eight modules, and the translation is done in eight modules.

This study is based on a statistical analysis of the text format in order to translate the ambiguous text into a specific meaning.

The study also used ANN dictionary for translation, translates similar meanings of the ambiguous word and gives each meaning a number, where the translation is in digital form, and the numbers are decoded into text.

The study focused on contextual text analysis as in previous studies .

11- Problems and Approaches to Translation with Special Reference to Arabic

Rami Al-Hamdalla (Associate Professor King Saud Univ, Vol 10, Lang & Trans)

Summary of the study:

The study deals with translation from the following aspects: Definition of translation, difference between translator and interpreter, relationship of translation by teaching English as a foreign language, information required for translation, difficulties in translation.

The study contains practical examples and problems faced by students of translation trainees at the university level.

The study discussed the general problems of translation as follows:

a- The analyzed parts did not jibe with the synthesized whole.
b- Some forms are found in one language but not in another language.
c- The grammatical value of inflected words is a problem.
d- There are words with a number of meaning.

12- Translating Names and Technical Terms in Arabic Text

Bonnie Glover Stalls and Kevin Knight USC Information Science Institute Marina del Rey, CA 90292

Summary of the study:

The study concerned the translation of names and terminology in the Arabic language. Specifically the Arabic letter where the letter translates in different letters in other languages .

For example, كريس (**krys**) comes out correctly as (**Chris**) and (**Kris**) but also, incorrectly, as **Grace**. The source of the **G, K** correspondence in the training data in the English name AE **L AH G Z AE N D ER Alexander,** which is الكسندر **Lksnd**r in our training corpus. A voiced fricative **G** is available in Arabic, which in other contexts corresponds to the English voiced **G**. English **X** is perceived to correspond to Arabic **Ks.**

Results and Discussion:

The ambiguous word has more than one meaning. Most previous studies discussed the ambiguity in translation and focused on text analysis (form), the analysis of the form may be statistically or analytically dependent on the description of the form by text, which is called the word tendency. Some studies concerned translating ambiguous words as symbols and numbers to be decoded into text.

The researcher proposes to divide the text in general into sections (scientific, mathematical, news, economic, medical), So that the text is inserted in any of the previous sections according to the analysis, ambiguous words are translated into text. But the statistical analysis must be activated because the text types are many and different, the inference feature can be used in artificial intelligence to translate ambiguous words.

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.

