The Thirteenth Conference on Language Engineering (ESOLEC'2013)

December 11-12, 2013

Organized by

Egyptian Society of Language Engineering (ESOLE)

Under the Auspices of

PROF. DR. HUSSEIN EISSA
President of Ain Shams University

PROF. DR. SHERIF HAMMAD
Dean, Faculty of Engineering, Ain Shams University

CONFERENCE CHAIRPERSON
PROF. DR. M. A. R. GHONAIMY

CONFERENCE COCHAIRPERSON
PROF. DR. SALWA ELRAMLY

Faculty of Engineering –Ain Shams University
Cairo, Egypt

http://esole-eg.org

Conference Chairman:

Prof. Dr. M. A. R. Ghonaimy

Technical Program Committee:

Prof. Taghrid Anber, **Egypt**
Prof. I. Abdel Ghaffar, **Egypt**
Prof. M. Ghaly, **Egypt**
Prof. M. Z. Abdel Mageed, **Egypt**
Prof. Khalid Choukri, ELDA, **France**
Prof. Nadia Hegazy, **Egypt**
Prof. Christopher Ciri, LDC, **U.S.A**
Prof. Mona T. Diab, Stanford U., **U.S.A**
Prof. Ayman ElDessouki, **Egypt**
Prof. Afaf AbdelFattah, **Egypt**
Prof. Y. ElGamal**, Egypt**
Prof. M. Elhamalaway**, Egypt**
Prof. S. Elramly, **Egypt**
Prof. H. Elshishiny**, Egypt**
Prof. A. A. Fahmy**, Egypt**
Prof. I. Farag, **Egypt**
Prof. Magdi Fikry, **Egypt**
Prof. Wafa Kamel, **Egypt**
Prof. S. Krauwer, **Netherlands**
Prof. Bente Maegaard, CST, **Denmark**
Prof. A. H. Moussa, **Egypt**
Prof. M. Nagy**, Egypt**
Prof. A. Rafae, **Egypt**
Prof. Mohsen Rashwan**, Egypt**
Prof. H. I. Shaheen**, Egypt**
Prof. S. I. Shaheen**, Egypt**
Prof. Hassanin M. AL-Barhamtoshy**, Egypt**
Prof. M. F. Tolba, **Egypt**
Dr. Tarik F. Himdi**, Saudi Arabia**

Organizing Committee

Prof. I. Farag
Prof. S. Elramly
Prof. H. Shahein
Dr. Mona Zakaria
Prof. Hany Kamal
Prof. M. Z. Abdelmegeed
Dr. A. Passant Elkafrawy
Dr. Bassant A. Hamid

Conference Secretary General

Prof. Dr. Salwa Elramly

Conference Sponsors

The Thirteenth Conference on Language Engineering
Final Program

Wednesday 11 **December 2013**

9.00 - 10.00 Registration

10.00 - 10.30 Opening Session

10.30 - 11.30 **Session 1: Invited Paper 1: Computational Linguistics**
Chairman**:** Prof. Dr. M. Adeeb Riad Ghonaimy

فروع الانسانيات الجدد
د. نبيل على
خبير اللغويات الحاسوبية

11.30 - 12.00 Coffee break

12.00 - 12.30 **Session 2** : **Invited Paper 2: Computational Linguistics**
Chairman : Prof. Dr. Ibrahim Farag

أحكام تنافر صوتى الفعل الثلاثى المضاعف: دراسة لغوية حاسوبية
أ.د/ وفاء كامل
كلية الآداب- جامعة القاهرة

12.30 - 14.00 **Session 3**: **Machine Translation**
Chairman: Prof. Dr. M. Z. Abdelmegeed

1.LILY: Language-to-Interlanguage-to-Language System Based on UNL
Sameh Alansary[*], Magdi Nagy[**]
[*]*Department of phonetics and Linguistics, Faculty of Arts, Alexandria University, Alexandria, Egypt.*
[**]*Computer and System Engineering Department, Faculty of Engineering, Alexandria University, Alexandria, Egypt*

2.MUHIT: A Multilingual Lexical Database
Sameh Alansary
Department of phonetics and Linguistics, Faculty of Arts, Alexandria University, Alexandria, Egypt.

14.00 - 15.00 Lunch

15.00 - 16.00 **Session 4: Room A: Language Analysis for Comprehension**
Chairman**:** Prof. Dr. M. Wafaa Kamel

1. Bel-Arabi Advanced Arabic Dependency Structure Extractor
Michael N. Nawar, Mahmoud N. Mahmoud
Computer Engineering Department, Faculty of Engineering, Cairo University Gamaet El Qahera St., Giza 12613, Egypt

2. Sentiment Analysis Improvement using the transformation of colloquial text to standard Arabic

Fatma El-zahraa El-taher, Alaa El-Dine Ali Hamouda, Salah Ramdan

Computer and System Department, Faculty of Engineering, Al-Azhar University, Nasr City, Cairo, Egypt.

3. Graph Reduction in Abstractive Text Summarization

Marwa Mahmoud, Ibrahim Fathy, Mostafa Aref

Department of Computer Science, Faculty of Computer and Information Sciences,Ain-Shams University, Cairo, Egypt.

16.00 - 16.30 **Session 5: Room A: Arabic Computational Corpus**

Chairman**:** Prof. Dr. M. Wafaa Kamel

ترميز الظواهر الدلالية في المعجم الحاسوبي العربي

وفاء كامل فايد، يوسف أبو عامر

كلية الآداب- جامعة القاهرة

15.00 - 16.30 **Session 6 : Room B : Speech Analysis and Synthesis**

Chairman: Prof. Dr. Mohsen Rashwan

1.The Creation of Emotional Effects for an Arabic Speech Synthesis System

Waleed M. Azmy, SherifAbdou, Mahmoud Shoman

Faculty of Computers and Information– Information Technology Department, Cairo Universiy, Cairo, Egypt

2. Performance of Different Speech Coders Over WiMAX and LTE

N. S. Abdelkader, N. A. Rasmy, H. H. Mourad

Electronics and Communication Eng. Department, Faculty of Engineering, Cairo university, Cairo, Egypt

3. Speech Compression Using Wavelet Packets Tree Nodes LPC Encoding and Best Tree Encoding (BTE) Features

Amr M. Gody[*], S. A. S. Abdelwahab[**], Tamer M. Baraket[*], M. Y. Mohamed[*]

[*]*Electrical Engineering Department, Faculty of Engineering, Fayoum University,* Egypt

[**]*Engineering Department, Nuclear Research Center, Atomic Energy Authority, Egypt*

Thursday 12 December 2013

10.00 - 12.00 **Session 7: Room A: Natural Language Processing (Students Session)**

Chairman : Prof. Dr. Hassanin M. Al-Barhamtoshy

1. Towards Building Automatic Scoring System for Assessing

Free-Constructed Responses
Nihal Al-Nazli, Sameh AlAnsary
Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University, ElShatby, Alexandria, Egypt

2. Towards Building a Rule-Based Dependency Parser for Modern Standard Arabic
Rehab Arafat, SamehAlAnsary
Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University, ElShatby, Alexandria, Egypt

3. A Syntactic Parser for Arabic Verbal Phrases Based on X-Bar Theory
Omnia Zayan, Sameh Al-Ansary
**Phonetic and Linguistics Department, Faculty of Arts, University of Alexandria, ElShatby, Alexandria, Egypt*

4.Semantic and Associative Relations in the Mental Lexicon: Evidences from Semantic Priming
Noha Fathy[*], Sami Boudella[**], Sameh Alansary[*]
[*]*Phonetics and Linguistics Department, Faculty of Arts-Alexandria University*
[**]*Faculty of Humanities and Social Sciences, United Arab of Emirates University*

5. Towards Arabic Named Entity Recognition Tool
Nouran Khallaf, Sameh Alansary
Phonetics Department, Faculty of Arts, Alexandria University
El-Shatby, Alexandria, Egypt

12.00 - 12.30 Coffee Break

12.30 - 13.30 **<u>Session 8</u>**: **Room A**: **Invited Paper 3: Language Engineering Frameworks and Methodologies**
Chairman : Prof. Dr. M. Fahmy Tolba
Real-World Natural Language Processing Based Solutions
Dr. Ossama Emam
IBM Cairo Human Language Technologies Group, IBM Egypt.

13.30 - 15.00 **<u>Session 9</u>: Room A: Natural Language Processing for Information Retrieval**
Chairman: Prof. Dr. Younis El Hamalawy

1. Content-Based Recommendation System Using Search Engine
Waleed M.Azmy[*], Ossama Emam[**]

*Information Technology Department, Faculty of computers and information, Cairo University, Giza, Egypt
** Human Language Technologies Group - IBM Cairo Technology Development Center, Giza, Egypt

2. Effective Mining and Visualizing for XML Semantic Structured Text

Z. T. Fayed, Tarek. M. Mahmoud**, M. M. Abdallah***
**Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt*
***Computer Science Department, Faculty of Science, Minia University, El-Minia, Egypt.*

3. Integrating CRF with GA to Build a Semi-Supervised Arabic Named Entity Recognition System

Noha Ahmed Saad*, Aly Farghaly**, Aly Aly Fahmy*
**Faculty of Computers and Information, Cairo University, Giza, Egypt*
***Text Group, Oracle USA, Redwood Shores, CA*

13.30 - 15.00 **Session 10: Room B: Optical Character Recognition**

Chairman: Prof. Dr. M. Fahmy Tolba

1. A Proposed Model for Standard Arabic Sign Language Recognition Using Artificial Neural Network

A. Samir Elons*, Magdy Aboul-ela**, M. F. Tolba*
** Scientific Computing Department- Faculty of Computers and Information Sciences- Ain Shams University, Cairo, Egypt*
*** Sadat Academy, Cairo, Egypt*

2. Identification Card Recognition Based on Arabic OCR System

Amira Abdel-Kareem, Ashraf Hussein, Esraa Shokry, Ola Alaa El-Din, Mohsen A. Rashwan, Hassanin M. Al-Barhamtoshy
Electronics and Communication Department, Faculty of Engineering, Cairo University

3. Using Word Based Features for Word Clustering

Farhan M. A. Nashwan, Mohsen A. A. Rashwan
Department of Electronics and Communications, Faculty of Engineering, Cairo University, Egypt.

4.An Advanced System for Arabic Name Entities Recognition

Wasim M. Abdulwasea, Dr. SherifAbdou, Hassanin Al-Barhamtoshy
Information Technology, Faculty of Computers and Information, Cairo University

15.00 - 16.00 Lunch

16.00 - 16.30 Closing session

الجمعية المصرية لهندسة اللغة

THE EGYPTIAN SOCIETY
OF LANGUAGE ENGINEERING

أعضاء الجمعية من المؤسسات

١ -مركز نظم المعلومات – كلية الهندسة – جامعة عين شمس
٢ -معهد الدراسات والبحوث الإحصائية – جامعة القاهرة
٣ -مركز الحساب العلمى – جامعة عين شمس
٤ -الأكاديمية العربية للعلوم والتكنولوجيا والنقل البحرى
٥ -أكاديمية أخبار اليوم
٦ -معهد بحوث الإلكترونيات
٧ -معهد تكنولوجيا المعلومات
٨ -مكتبة الإسكندرية
٩ -المعهد القومى للاتصالات (NTI)
١٠ -الشركة الهندسية لتطوير نظم الحاسبات (RDI)
١١ -الهيئة القومية للاستشعار من بعد و علوم الفضاء
١٢ -كلية الحاسبات و المعلومات جامعة قناة السويس
١٣ -دار التأصيل للبحث و الترجمة

أهداف الجمعية

١ -الاهتمام بمجال هندسة اللغويات مع التركيز على اللغة العربية بصفتها لغتنا القومية والتركيز على قواعد البيانات المعجمية وصرفها ونحوها ودلالتها بهدف الوصول إلى أنظمة ألية لترجمة النصوص من اللغات الأجنبية إلى اللغة العربية والعكس ، وكذلك معالجة اللغة المنطوقة والتعرف عليها وتوليدها، ومعالجة الأنماط مع التركيز على اللغة المكتوبة بهدف إدخالها إلى الأجهزة الرقمية.
٢ -متابعة التطور فى العلوم والمجالات المختصة بهندسة اللغة
٣ -التعاون مع الجمعيات العلمية المماثلة على المستوى المحلى والقومى والعالمى.
٤ -إنشاء قواعد بيانات عن البحوث التى سبق نشرها والنتائج التى تم التوصل إليها فى مجال هندسة اللغة بالإضافة إلى المراجع التى يمكن الرجوع إليها سواء فى اللغة العربية أو اللغات الأخرى.
٥ -إنشاء مجلة علمية دورية للجمعية ذات مستوى عال لنشر البحوث الخاصة بهندسة اللغة و كذلك بعض النشرات الدورية الإعلامية الأخرى بعد موافقة الجهات المختصة.
٦ -عقد ندوات لرفع الوعى فى مجال هندسة اللغة
٧ -تنظيم دورات تدريبية يستعان فيها بالمتخصصين وتتاح لكل من يهمه الموضوع. وذلك من أجل تحسين أداء المشتغلين فى البحث لخلق لغة مشتركة للتفاهم بين الأعضاء
٨ -إنشاء مكتبة تتاح للمهتمين بالموضوع تشمل المراجع وأدوات البحث من برامج وخلافه.

٩ -خلق مجال للتعاون وتبادل المعلومات وذلك عن طريق تهيئة الفرصة لعمل بحوث مشتركة بين المشتغلين فى نفس الموضوعات.
١٠ -تقييم المنتجات التجارية أو البحثية والتى تتعرض لعملية ميكنة اللغة.
١١ -رصد الجوائز التشجيعية للجهود المتميزة فى مجالات هندسة اللغة.
١٢ -إنشاء فروع للجمعية فى المحافظات.

المؤتمر الثالث عشر لهندسة اللغة

١١-١٢ ديسمبر٢٠١٣

جمهورية مصر العربية-القاهرة

ينظم المؤتمر

الجمعية المصرية لهندسة اللغة

تحت رعاية

الأستاذ الدكتور/ حسين عيسى
رئيس جامعة عين شمس

الأستاذ الدكتور/ شريف حماد
عميد كلية الهندسة - جامعة عين شمس

رئيس المؤتمر
الأستاذ الدكتور/ محمد أديب رياض غنيمى

مقرر المؤتمر
الأستاذ الدكتور / سلوى حسين الرملى
كلية الهندسة - جامعة عين شمس

مكان عقد المؤتمر : كلية الهندسة – جامعة عين شمس

http: // www.esole-eg.org

Table of Contents

X. Language Generation

Identification Card Recognition Based on Arabic OCR System

Amira Abdel-Kareem[1], Ashraf Hussein[2], Esraa Shokry[3], OlaAlaa El-Din[4]

Mohsen. A. Rashwan[5], Hassanin M. Al-Barhamtoshy[6]

Electronics and Communication Department, Faculty of Engineering-Cairo University

amira_shaban22@yahoo.com

h_ashraf16@hotmail.com

engesraa_2013@hotmail.com

olaalaa_2013@yahoo.com

mrashwan@rdi-eg.com

hassanin@kau.edu.sa

***Abstract*— Optical character recognition of Arabic language is a field of research that is socially very relevant and challenging, the social relevance lies on the fact that OCR is very important for many applications that need character recognition of image. Our system is Egyptian ID cards reader system which extracts important data from the ID card image, recognizes them and translates them into editable text on computer so they can be edited and saved, and then the system can verify between a testing ID card and the database saved before. This paper extensively reviews the base line-based segmentation and DCT based feature extractor approaches used forbuilding this special Arabic OCR system. It also reports the experimental results obtained so far showing the reliability of our system. Finally, the system works fast on the scientific Matlab program as it needs about 16 seconds in average to process one ID card, and the system is expected to do better performance when transferring it from the academic phase to the product phase.**

1 INTRODUCTION

Humans recognize characters easily and they repeat the character recognition process thousands of times every day as they read papers or books. Though, after many years of serious investigation and research, the ultimate goal of developing an optical character recognition (OCR) system with the same interpretation capabilities as humans still remains unachieved. One of the main objectives of an OCR is to reach a 5 characters/second speed with a 99.9% recognition rate, with no errors.

The OCR is the mechanical or electronic conversion of scanned images of typewritten, printed text, or handwritten, into machine-encoded text. OCR allows the machine automatically to recognize characters in an image and translate them into computer textual format by applying machine learning mechanism.

Therefore, development of the OCR systems is a very significant field of research in pattern recognition. Different OCR engines allow the machine to automatically recognize characters in an image and translate them into computer textual format by applying machine learning mechanism. It improves human-machine interaction and is widely used in many areas. Here's a general (OCR) DFD diagram, as shown in Fig. 1.

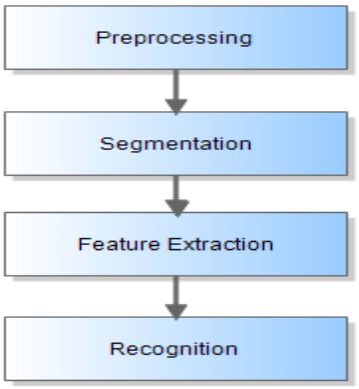


Figure 1: OCR DFD Diagram

Recently, identification cards recognition systems became very important due to the automatic processes of checking and storing card's data achieved by them in many applications such as election systems and vital governmental installations and makes this operation easy and fast using optical solutions.
So from the realization of the importance of such applications and its impact on the society, this paper shows developing an Arabic OCR system dedicated for ID card recognition.

An over view of the system will be presentedin section 2, Arabic script characteristics and Arabic OCR challenges are discussed in section 3. The proposed OCR algorithm will be discussed in section 4.Experimental work and results are resulted in section 5. Finally, conclusion and future work will be discussed in section 6.

2 SYSTEMOVERVIEW

The data acquisition system consists of a simple wooden structure scanning model. This model has a special place for the mobile and another one for the ID card, the image is captured with a 5 megapixel resolution camera of an Android mobile and the captured images' resolution is 72 dpi.

Using this simple scanning process, images of front and back sides of the ID card are captured by mobile, and then the captured images are sent automatically to the computer by Android service installed in such mobile. The captured image is then passed and processed via the system's software (Created Matlab program) and the results will be outputted through simple Graphical User Interface (GUI) which contains all the recognized available ID data, see Fig. 2.

Figure 2: System Overview

3 ARABIC SCRIPT CHARACTERISTICS AND OCR CHALLENGES

Arabic is the official language of over twenty Arab countries which stretch from Morocco to Iraq, it is the religious language of all Muslims of more than one billion Muslims spread all over the world and it is the language of the Quran (the sacred book of Islam). Arabic language is a Semitic language and most of its words are built up from roots by following specific morphological grammatical rules by employing affixes processing (infixes, prefixes and suffixes). Classical Arabic language is widely used in around fourteen centuries ago.

Arabic is a popular script and cursive nature language. More than one billion Arabic script users are estimated in the world. Due to the cursive nature of Arabic script, the development of Arabic OCR systems involves many technical troubles, particularly in the segmentation stage. Although many researchers are studying solutions to solve such troubles very little progress has been made.

A. Arabic Script Characteristics

Arabic is written from right to left, and so recognition process should occurred from right to left. Arabic character set includes 28 letters; each Arabic letter has 2-4 different forms which depend on its position in the word. Fifteen of the 28 letters have dots and the other 13 are without dots. Dots are above and below the Arabic letters, it plays a major role in discriminating some characters from each similar, that differ only by the number or location of dots; e.g.letters (ب- ت- ث-

ي- ن). There are four characters which may take the secondary character "Hamzahء", those are "Alif ,"أإ"Waw ؤ" , "Yaaئ" and "Kaf" ك. Six Arabic letters can be connected from the right side only: dal (د), raa (ر), waw(و), alef (ا), thal (ذ), and zay(ز). Whilethe other 22 letters can be connected from both sides. These six characters have only two forms, the stand-alone form and the final form.

Arabic letters do not have rigid width or fixed size, even in printed form. The shape of the letter is influenced by its position in the word. Whereas the rest of the characters scan appear in any of four forms: the initial, the middle, the final, and the stand-alone form. Consequently, an Arabic word may consist of one or more sub-words. A sub-word can be defined as the basic stand-alone pictorial block of the Arabic writing.
Any optical character recognition of Arabic characters should treat the sub-word as the basic block for processing whatever the method it uses for preprocessing, segmentation, recognition, or classification. This is because each sub-word is separated from other sub-word by a space. Although spaces between sub words are usually shorter than those between successive words, still they are surrounded by space. A word may contain one or more sub-words. Some of these sub-words may even consist of a single character in its stand-alone form.
Hence, their recognition does not need segmentation. Shape of the letter in the text differs according to the location of the character in the sub-word, i.e. a character at the end of sub-word, has exactly the same shape when it comes at the end of a full word.

The Arabic character set is shown in Table I that illustrates the variation of the Arabic characters' shape depending on their positions in the word,

TABLE I
THE DIFFERENT FORMS OF ARABIC ALPHABETS

Character Name		Isolated	Initial	Middle	Final
Alif	ألف	ا	ا	ـا	ـا
Ba'	باء	ب	بـ	ـبـ	ـب
Ta'	تاء	ت	تـ	ـتـ	ـت
Tha'	ثاء	ث	ثـ	ـثـ	ـث
Jeem	جيم	ج	جـ	ـجـ	ـج
H'a'	حاء	ح	حـ	ـحـ	ـح
Kha'	خاء	خ	خـ	ـخـ	ـخ
Dal	دال	د	د	ـد	ـد
Thal	ذال	ذ	ذ	ـذ	ـذ
Rai	راي	ر	ر	ـر	ـر
Zai	زاي	ز	ز	ـز	ـز
Seen	سين	س	سـ	ـسـ	ـس
Sheen	شين	ش	شـ	ـشـ	ـش
Sad	صاد	ص	صـ	ـصـ	ـص
Dhad	ضاد	ض	ضـ	ـضـ	ـض
Tta'	طاء	ط	طـ	ـطـ	ـط
Dha'	ظاء	ظ	ظـ	ـظـ	ـظ
A'in	عين	ع	عـ	ـعـ	ـع
Ghain	غين	غ	غـ	ـغـ	ـغ
Fa'	فاء	ف	فـ	ـفـ	ـف
Qaf	قاف	ق	قـ	ـقـ	ـق
Kaf	كاف	ك	كـ	ـكـ	ـك
Lam	لام	ل	لـ	ـلـ	ـل
Meem	ميم	م	مـ	ـمـ	ـم
Noon	نون	ن	نـ	ـنـ	ـن
Ha'	هاء	ه	هـ	ـهـ	ـه
Waw	واو	و	و	ـو	ـو
Ya'	ياء	ي	يـ	ـيـ	ـي

B. Arabic OCR Challenges

Although working on a standard ID has advantages like, the font type is fixed (simplified Arabic font) and approximately similar font sizes, Arabic words may horizontally overlap and characters may stack on others. These introduce problems for both the word and the character segmentations. At this stage, it is not hard to understand that segmentation is a crucial step in the development of an Arabic OCR system. The main difficulty associated in Cursive text recognition is the segmentation of words to characters. And here's an introduction to the main challenges in Arabic OCR system,

1) Connectivity Challenge: As illustrated before that the cursive phenomenon of Arabic language text introduces problems for both the word and the character segmentations. The main difficulty associated in cursive text recognition is the segmentation. Accordingly, the segmentation is a critical step in the development of an Arabic OCR system. Fig.3.illustrates Arabic naming script showing connectivity.

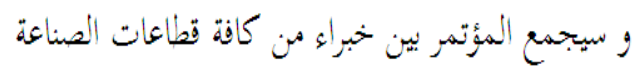


Figure 3: Example of the Arabic Script

2) The Dotted Challenge: Dotting is extensively used to differentiate characters sharing similar graphemes. Fig.4. shows some example sets of dotting-differentiated graphemes, it is apparent that the digital differences between the members of the same set are small. Whether the dots are eliminated before the recognition process, or recognition features are extracted from the dotted script, dotting is a significant source of confusion, hence recognition errors in Arabic OCR systems especially when run on noisy documents; e.g. those reproduced by photocopiers.

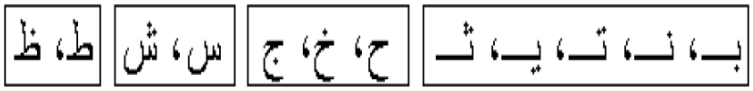


Figure 4: Example sets of dotting-differentiated graphemes

*3) Multiple Grapheme Cases Challenge :*Due to the mandatory connectivity in Arabic orthography; the same grapheme representing the same character can have multiple variants according to its relative position within the Arabic word segment(Starting, Middle, Ending, Separate) as exemplified by the 4 variants of the Arabic character "ع" shown in bold in Fig. 5.

Figure 5: Grapheme "ع" in its 4 Positions

4) Character's Size Variation Challenge: Different Arabic graphemes do not have a fixed height or a fixed width. Moreover, neither the different nominal sizes of the same font scale linearly with their actual line heights, nor the different fonts with the same nominal size have a fixed line height.

At this point, many of challenges have been illustrated in the Arabic script characteristics, in the next section;the algorithm of the proposed OCR system will be introduced.

4 THE PROPOSED ARABIC OCR ALGORITHM

This section discusses the algorithm of the proposed OCR system in details,Fig.6shows simple block diagram of theproposed system, next subsections will explainthe related algorithm andthe main steps,

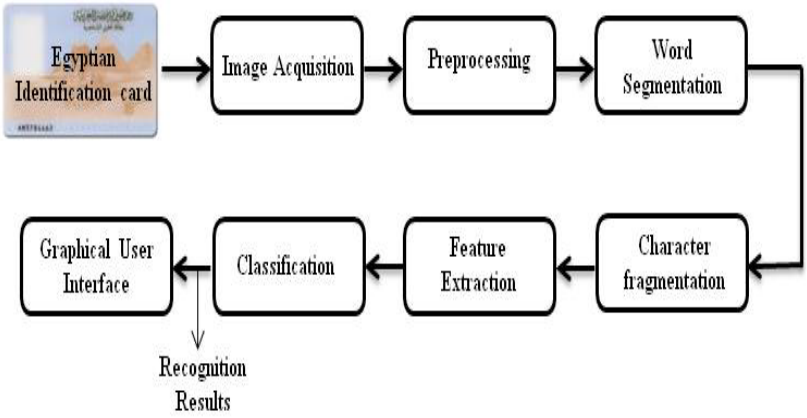


Figure 6: The Proposed Arabic OCR Block Diagram

A. OCR Algorithm

The proposed OCR system consists of 4 main steps which will be introduced within the few following subsections,

*1) Image Acquisition:*As illustrated before, we capture an image of the ID card with an android mobile with 5 Megapixel resolution camera, the captured image is 72 ppi (pixel per inch).

*2)Preprocessing and Edge Detection:*Preprocessing is a very important and main step in image processing as scanned images are usually displayed in gray scale or color and also they suffer from noise, varying spaces between letters, varying space between lines and other image problems.
Noise removal and edge detection are the two most important steps in processing of any digital images to improve the information in the picture so that it can be easily understood by man and to make it suitable and readable for any machine working on those images. So some preprocessing steps are performed on the image,

Firstly, in the step of thresholding and binarization, the gray image is converted to a 'binary' image. We mean by binary that the image is presented by black and white pixels only. This helps us to work more efficiently on the image than in the gray or the colored form. Binarization is achieved through the process of thresholding in which a 'threshold' value is chosen and any pixel with a value greater (or less) than this value is converted to a text (orbackground) pixel. That is, its value is made either 0 or 255.

Noise is the unnecessary information that exists in the image which may have been inadvertently introduced. This may be because of inefficient input devices used. To remove the noise which may affect the performance of the system, filters are used. So firstly, the "median filter" is used as an example of a non-linear spatial filter, recall that the median of a set is the middle value when they are sorted which is used to remove the salt and pepper noise from the card. Finally, remove the undesired borders by clipping it. Edge detection is a very important step because the edge is one the important and basic features of an image. If the edges of an image are identified accurately, some basic properties such as area, perimeter and shape can be measured. Edge detection can play a signification role in different fields such as computer vision, pattern recognition, image segmentation, remote sensing and medical image analysis.Many classical edge detectors have been developed over time. However classical edge detectors usually fail to handle images with strong noise.

Mathematical morphology (MM) is a new mathematical theory which can be used to process and analyze images. In the MM theory, images are treated as sets, and morphological transformations which derived from Minkowski addition and subtraction are defined to extract features of images.
Morphological filters are nonlinear signal transformations because image is probed by a structuring element which interacts with the image in order to extract useful information about the geometrical structure of the image and achieve the goal of preserving thin features while removing noise. So, this algorithm will be applied in the stage of card recognition.
The structuring element is considered to be the building block of the dilation (expanding features) and erosion (shrinking features) processes and by the way it is the mainly constructing element of the morphological image processing, it is represented by matrix of 0s and 1s, sometimes it is convenient to show only the 1's. The origin of the structuring element must be identified. It could have any shape but its size must be smaller than the original image's size. Fig.7. shows some shapes of structuring elements.

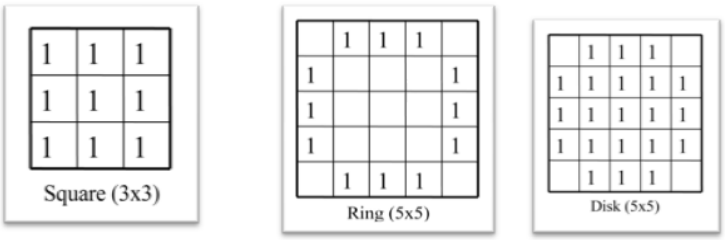


Figure 7: Some shapes of structuring elements

In the proposed system, the rectangle and line (with 0 & 90 degrees) structuring elementsare used as there were more suitable for our dilation and erosion operations, see Fig. 8.

Figure 8: Line and rectangle structuring elements

Dilation is an operation that grows objects in binary images controlled by a shape referred to as a structuring element. Dilation of original image (A) by structuring element (B) means that for each point x belongs to B, we translate A by those coordinates then we take union of these translations, as shown in the following equation,

$$A \oplus B = \{z|(\hat{B})_z \cap A \neq \emptyset\}$$

The above equation denotes the dilation of A by B is the set consisting of all the structuring element origin locations where the reflected and translated B overlaps at least some portion of A. Such translation of the structuring element in dilation is similar to the mechanics of spatial convolution.

Erosion is an operation that thins or shrinks objects in binary images controlled by a shape referred to as a structuring element. Erosion of original image (A) by structuring element (B) means that the output image has a value of '1' at each location of the origin of (B), such that the element only overlaps 1-valued pixels of (A).

$$A \ominus B = \{z|(B)_z \cap A^c \neq \emptyset\}$$

The above equation denotes the erosion of A by B is the set of all structuring element origin locations where the translated B has no overlap with the background of A.

3) Feature Extraction Module: Features extraction is one of the most important factors in achieving high recognition performance in pattern recognition systems. It has been defined as the process of extracting information that is mostly useful for the purpose of classification from the raw data, as it involves simplifying the amount of resources required to describe a large set of data accurately, so it is considered to be a special form of dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant, the input data can be transformed into a reduced representation set of features (named feature vector). So transforming the input data into set of features is called the feature extraction.

Discrete Cosine Transform (DCT) is one of the most efficient and popular techniques used in feature extraction. In particular, a DCT is a Fourier-related transform similar to Discrete Fourier Transform (DFT), but DCT is more efficient in data reduction and energy compaction as it stores most of image details in few coefficients as shown in Fig.9.

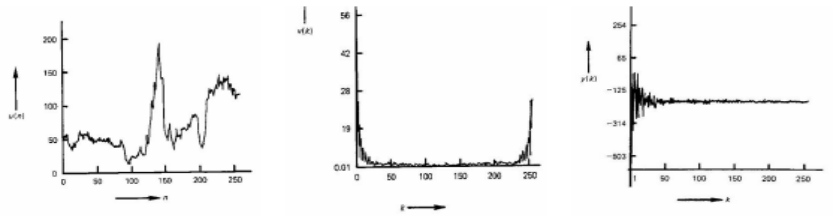


Figure 9: Energy distributions in different transforms

The proposed solution uses ideal mathematical tool of the Discrete Cosine Transform DCT. Such DCT based feature extractor is often used in signal and image processing especially for lossy data compression, because it is capable of packing the energy of spatial sequence into few coefficients as possible so it has strong energy compaction property. So a larger number of coefficients get wiped and great bit savings for the same loss.

Two Dimensional -DCT is applied to the whole image then most of signal information tends to be concentrated in a few low-frequency components and approximately most of the important data and details of the image are then allocated at the upper left corner of the image as shown in Fig.10.

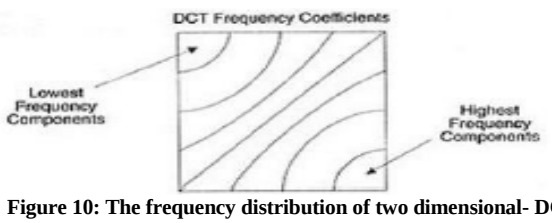


Figure 10: The frequency distribution of two dimensional- DCT

Eye is most sensitive to low frequency components (upper left corner), so Zig-Zag scanning methodis used to group low frequency coefficients in top of a vector with a certain technique as shown in Fig. 11.
After applying Zigzag scan, a vector of 20 elements is created for each model in the training set; this vector is the feature vector that will be used later in the next stage of classification and decision making.

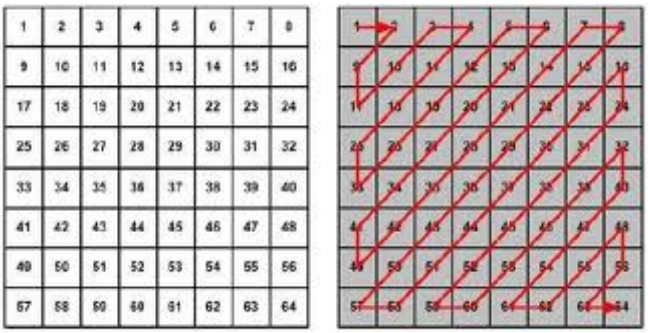


Figure 11: Zig - Zag Scan

4) Classification and Decision Making: The main role of the classifier is to compare the feature vector of each block of data segmented from ID card with the previously built model of the training set, and then provide the system with information about the nearest neighbor to this block of data and the Euclidean distance between them. The Euclidean distance between any 2 vectors (q and p- each one of them has (n) elements) can be calculated according to the following formula:

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n} (q(i) - p(i))^2}$$

Where, "q" and "p" are two feature vectors with size (n) equals 20 elements.

B. ID Recognition Algorithm Steps

First, the preprocessing steps are applied on the whole ID card except the erosion step, after performing dialation or filling operation theresultwill be displayed as shown in Fig. 12.

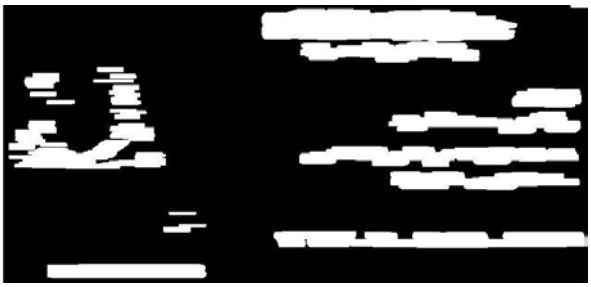


Figure 12: ID card after Dilation Process

Then the wanted data is detected and then the areas of the data are segmented with rectangular frames, each type of data is sent to its specified function, as shown in Fig. 13.

Figure 13: The Detected Text

There are two types of algorithms depending on the data types, an algorithm for numbers like ID National number, release date and expiry date of the ID cards and the other for words which is specified for all other types of data.

The functions are (First Name & Last Name – Address – ID National number – Job – Job Address – Religion & Type & Marital status – Husband name – Expiry date& Release date).A training data set containing all forms of the Arabic language alphabets, numbers from (0-9) and some characters like (/ , -), is passed to the words function and number's function respectively, and also

1) Numbers Function's Algorithm: A training data set containing all numbers from (0-9) is passed to this function, and also some characters like (/, -).
Firstly, we get the feature vectors for the numbers (0-9) as a data set, and thenwe get the feature vectors for the unknown national ID card number which we want to detect its 14 numbers, Fig.14 shows ID card number.

Figure 14: ID card number

The following steps will be used to perform preprocessing:
Thresholding and binarization: In this step, the RGB ID number image is converted to a 'gray' image, then to 'binary' one.
Noise filtering: the "median filter" is used as an example of a non-linear spatial filter to remove the salt &pepper noise from our card.
Mathematical morphology (MM): which consists of two main basic operations Dilation and Erosion, the line (with 0 & 90 degrees)structuringelement is more suitable for dilation and erosion operations. Fig. 15 shows ID card number after performing preprocessing steps.

Figure 15: ID card numberafter performing preprocessing steps

Segmentation: Then by using some functions, the ID number image is divided into 14 segments (regions). Then for each region (number) we get its properties (area, centriod, bounding box) and by using the bounding box vector which contains [X_{min}, Y_{min}, width, height] we could detect its boundaries and the image is segmented and ready for feature extraction stage Fig. 16 shows ID card numbers after segmentation.

Figure 16: Segmented ID card numbers

For each region which represents one number in the ID number, we get the feature vector and by using our classifier we get the location of the nearest number to database, also we get minimum distance between the region feature vector and the other feature vectors in database. According to this minimum Euclidean distance, we can take the right decision. Finally we have a vector containing the 14 numbers of the National ID Number.

2) Words Function's Algorithm: In these functions the training data set containing all forms of the Arabic language graphemes. The proposed algorithm for segmentation and detection of Arabic words is starting to get the feature vectors for each Arabic character from the database.Then three steps are applied to perform preprocessing: (1) Thresholding and binarization; (2) Noise filtering; and (3) Math morphology (Dilation and Erosion).

Deskewing to fix the wrongly rotated words: to remove the undesired rotation in the image, rotate the image with many angles, for each angle apply the horizontal projection profiles and find the highest peak, then choose rotation angle equal to the angle which has the highest peak and rotate the image by it. Deskewing mechanism is showing good performance with the fieldsthat have many words but showing bad performance with the single words, so for the fields of data that have just one word like first name field, the deskewing mechanism is deactivated, but for sentences like address or last name, it's useful to activate this mechanism.

Figure 17: The image after filling and deskewing

Figure 18: The image after erosion and deskewing

After the cut area is sent, words segmentation and separation is done by noticing the intra-spaces between words of the sentence, by experiment the number of zeros (space) between words is determined to be 25 zero minimum. Sentence image is vertically projected, by applying vertical projection on image, the vertical projection profile is defined as:
$P(j) = \sum image(i, j)$, where p(j) is the vertical projection of the image for column j and the image(i, j) is the pixel value at (i, j).
And by using this vertical projection, the spaces in the sentence can be located by applying a certain condition on the length of these spaces, so the location of spaces between words in the sentence can be easily recognized, the regions of words are detected and we got segmented words and the words are separated, see Fig.19.

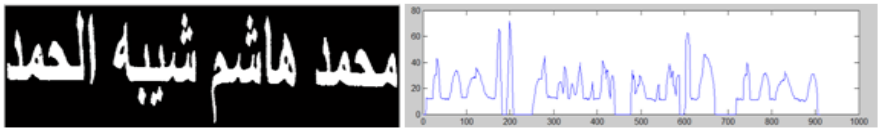


Figure 19: The words after segmentation

Detecting baseline is one of the main majorities in the preprocessing stepof Arabic OCR system, as it can be used for both skew normalization and segmenting the text into words or characters.

The baseline detection is very important in Arabic OCR, because it can be used to segment the Arabic text to characters and make the text ready for the feature extraction stage. Also baseline has been used by most of the OCR systems.

The horizontal projection method is used by the OCR researchers to detect Arabic baseline, and it works well with the printed text. This method detects the Arabic baseline by reducing the 2D of data to 1D based on the pixels of the text image, and the longest peak after projection will define the baseline range, as illustrated in Fig.20.

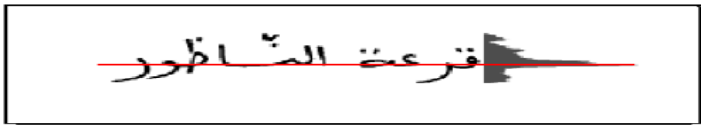


Figure 20: Base line detection by horizontal projection

The horizontal projection profile is defined as: $P(i)=\sum image(i, j)$,where p(i) is the horizontal projection of the image for row i,and the image(i , j) is the pixel value at (i , j).
Determine the baseline: that depends on the iteration with angle to detect Arabic baseline with the horizontal projection, the highest peak corresponding to our rotation angle is used to determine the baseline, and Fig.21 shows text after detecting the baseline.

Figure 21: Text after detecting the baseline

For fragmentation of words to characters, we apply vertical projection to the word to detect the beginning and ending of each character in the word as shown in Fig.22.

Figure 22: Character Fragmentation

The classification step starts with the feature vector of the unknown segmented character, by applying the proposed classifier using the Euclidian distance between the unknown character feature vector and all training feature vectors. According to the minimum distance we can know the nearest character and take the right decision.

3) Gender, Religion and Marital status Function's Algorithm: To increase the system accuracy, we make a separate function for the standard words in the national ID card like gender, religion, and marital status.
Firstly, we get the feature vectors for these standard words as a data base.For religion ('مسلم','مسلمة','مسيحي','مسيحية')' , For marital status ('أعزب',' انسة',' متزوج','متزوجة','مطلق','مطلقة',' ارمل',' ارملة') and For gender ('ذ كر',' انثي').
Then for the unknown word which we want to know, we apply two steps to perform preprocessing,

Thresholding and binarization: In this step, the RGB image is converted to a 'gray' image, then to 'binary' one.Noise filtering: we used the "median filter" as an example of a non-linear spatial filter to remove the salt &pepper noise from our card.Then we apply the mathematical morphology (MM): which consists of two main basic operations; Dilation and Erosion, we used the line (with 0 & 90 degrees) structuring element as there were more suitable for our dilation and erosion operations.

For the decision step, we get the feature vector for the unknown segmented area, by using the proposed classifier we get the minimum Euclidean distance between this feature vector and the feature vectors in the standard words database. According to this minimum Euclidean distance, the right decision can be taken.

5 EXPERIMENTAL WORKANDRESULTS

This section presents the results obtained, and discusses the limitations and problems which affect the accuracy of the system, and also discusses solutions for some of them.
The number of collected ID cards is 40, images are captured by a 5 megapixels camera which takes pictures with resolution 72 ppi, they are ensured to be taken in random and different environments as the changes in brightness or source of light or the homogeneity will affect the accuracy, the ideal situation of scanner solves this problem as it provides a homogeneous, fixed distribution of light for all parts of the ID card, for the proposed system, sunny environment provides this homogeneous distribution then it's chosen to be the default environment for the design, and as we'll see in the following sections that the change of environment is a reason of some bad accuracy, the 40 ID cards are divided into 10 ID cards for training phase and 30 ID cards for testing phase, Table II shows the accuracy of the system for the testing phase before doing corrections.

TABLE II
SYSTEM ACCURACY BEFORE CORRECTION

Phase	System Accuracy		
	Total Number	Correct Number	Percentage
Segmentation	374	360	96.26%
Numbers	1249	1246	99.76%
Words	726	675	92.97%

The average time of run for a single card reaches 16 seconds, and the maximum time of run fora single card is 22 seconds. The following sections will discuss some reasons that results in errors at segmentation and words recognition phases, and then will suggest some solutions to increase system's accuracy.

A. Problems of Segmentation Phase

Segmentation phase is the process of cutting out the important information from the ID card after finishing the preprocessing phase.In case of ID card the locations of data are almost fixed so it can be defined easily if segmented information is the first name, last name … etc., but in some ID cards there are some unusual situations causing errors in segmentation process. The following subsections will illustrate some of these problems noticed from our experimental testing.

*1) Wrong locations of information problem:*As previously illustrated, the location of information is almost fixed in different ID cards, so every segmented information can be definedfrom its location. But as displayed in Fig.23, the

information in the back of ID card in figure (b) is shifted obviously up than any other normal ID card and we can notice that by comparing the location of information in both cards.

This problemmakes the system detects the job address field in ID card in Fig. (b) in the field of job and the religion in the field of job address, and can't define the job, marital status and husband name information as shown in Fig.23-c.

(a) (b)

Job :	مدرسة [redacted] الابكانية
Job address :	معلم
Gender :	أنثى
Religion :	not found
Marital state :	not found
Husband name :	not found

(c)

Figure 23: Results of wrong locations of information

Another problem of wrong located information, cards which don't have expiry date information field for married females, husband name is shifted down and the system segments it as expiry date which causes an error of segmentation .Fig. 24 (a)shows an ID card with normal husband name location but Fig. (b). Shows a wrong location of husband name in other card.

(a) (b)

Figure 24: Wrong husband name location

2) Dilation Process problem : Dilation process is used to fill the spaces between the words, so the locations of text can be defined and then the system can define the type of this information, but in some cases, some information fields could be under dilated or over dilated, under dilation problem would cause partially successful segmentation, and over dilation problem is linking an unnecessary information to a necessary one, so system will be unable to detect the necessary information and considers it missing information.

3) Noisy Information problem: If there is noisy information in the ID card caused by an error in preprocessing or resulted from under dilated information as previously illustrated, this noisy information is detected as necessary information which causes missing of the needed information.

As illustrated in Table III, for about 374scanned cardsto be segmented, the system successfully segmented 360 fields of information and failed in 14, Table III shows the error rate of each problem previously illustrated.

TABLE III
ERROR RATE OF SEGMENTATION PHASE PROBLEM

	Wrong locations problem	Dilation process problem	Noisy information	Other
Error Rate	10	2	1	1

B. Problems of Words Recognition Phase

Words recognition process is the process of recognizing characters in an image and translating them into computer textual format by applying the recognition mechanism on the image,various errors happen in the experimental work, the following subsections discuss some common errors noticed in testing phase.

1) Baseline Detection problem: The first step of baseline detection process is getting the horizontal projection of word image, then defining the row that has maximum value in the projection then defining a range of rows around the row of the maximum value, but in some testing words, this range is not enough to detect the base line accurately, and this happens because of different size of words in the same field for different ID cards or because of theskewing of the image was not fixed at the deskewing process. Failing in detecting the baseline accurately would cause failing in characters segmentation as the characters get linked together as shown in Fig.25, this is a wrong recognized word due to wrong detection of baseline.

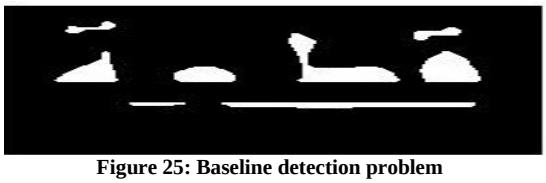


Figure 25: Baseline detection problem

Another reason of linking characters is the Arabic character 'ع ' when it appearsas the first character of the word, in some cases it sticks with the following character because of the very small space between them as can be seen in Fig. 26 , so the baseline detection process can't separate them accurately.

Figure 26: Stuck characters problem

2) Over and Under Segmentation problem: In character segmentation process, when words have different sizes than the default size in a certain field, this leads to wrong segmentation for some characters like the Arabic character 'س' for example, over-segmentation is when single character is segmented to more than one segment, and under-segmentation is when more than one character are considered as one character, Fig.27illustrates the two cases.

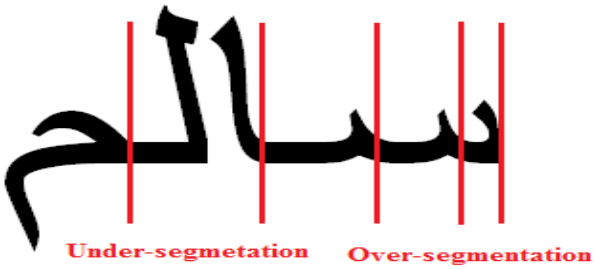


Figure 27: Over and under segmentation

3) Over Erosion: As previously mentioned that for the same field of information e.g. job field the words could differ in size for different ID cards, and as the erosion process is fixed for all words in the same field, over-erosion could happen to the smaller size words, it results in errors in the recognition process, one of the common problems caused by over-erosion is errors occurred in the word which has the Arabic character 'ء/hamza', over-erosion is affecting this character as the eroded 'hamza' is recognized by the system to be a dot. Such problem represents a word has over eroded 'hamza', the system recognized it as another Arabic character 'ـنـ/noon' in the recognition process. Fig.28 shows an over erosion example.

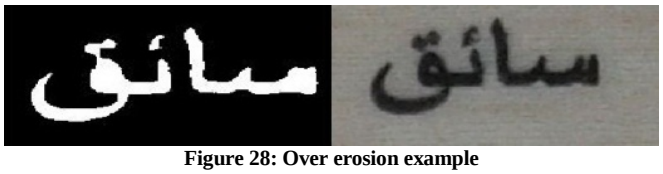


Figure 28: Over erosion example

4) Environment Effect on the System Accuracy: It's mentioned before that taking pictures for the ID cards in different environment's conditions affects the system accuracy, as the image of the ID card in a certain environment could be clearer than another image for the same ID card in different environment, and it should be considered that the testing ID cards' pictures are taken in different environments to test the capability of the system to deal with different circumstances, and it's shown that the more clearer picture with homogeneous light distribution the more accuracy is obtained. Fig.29 shows the response of the system to two pictures for the same ID card but in different environmental conditions.

Figure 29: Same ID card in two environmental conditions

C. Suggested Correction

As we represented the problem of wrong location of husband name information in most of ID cards of females which don't have expiry date information field, a correction for this problem is suggested that depends on linking the information fields with each other, by taking the advantage that the functions of gender, religion and marital status are perfectly recognized in all tested ID cards because they are depending on correlation method of recognition. So if the system detected that the ID card is for a married female, so the husband name should be found, and if it's not found the system checks the output of expiry date field recognition function, if the output is not logical enough, the system refuses this output and sends the information of expiry date field to the function of husband name recognition. Fig.30 shows the output of an ID card that has this problem before and after the correction.

Figure 30: Correction effect

This algorithm solved 4 errors of the 14 errors in the segmentation phase, and to ease following-up the algorithm, we'll write it in steps.Run the functions of gender and marital status recognition.If the card is for married female, check the husband name.If there is no information in husband name field, run the function of expiry date recognition.

If the output of expiry date function is not logical enough, refuse this output and send the information in expiry date field to the function of husband name recognition.If the output is logical, accept this output as expiry date and inform that you didn't find husband name.

D. Final results

The correction enhances the accuracy of segmentation phase as it solves most of the wrong location cases of husband name, the final results of testing 30 ID cards is shown in Table IV.

TABLE IV
SYSTEM ACCURACY AFTER CORRECTION

Phase	System accuracy (after correction)		
	Total number	Correct number	Percentage
Segmentation	374	364	97.33%
Numbers	1249	1246	99.76%
Words	743	688	92.60%

6 CONCLUSIONS AND FUTURE WORK

To summarize, the proposed Arabic OCR system is used to extract data from the image of Egyptian ID cards and then recognize these data and translate it into computer textual format. This system could be used in creating database for a lot of ID cards easily and fast or in verification between data of ID card and previously saved database. The evaluation of the presented system is excellent in data segmentation and number recognition and very good in recognizing characters due to the challenges facing the system in this phase.

As was discussed before, the proposed system is very helpful for business checking and saving personal ID’s information. Also, it saves valuable time needed to type these data manually and also gives a very good recognition accuracy, which makes the system reliable enough to be applicable in governmental authorities or even companies. There are many applications that could need to use this system, like election system, Wallet services, Banking card, security services …etc.

The proposed system solution could be enhanced in several ways, the first way is to make better design for the process of taking shots by camera to provide the system with fixed source of light and with fixed intensity in any environment, or using a scanner provided by motor to satisfy fast processes, this way of development will end most of the common problems facing our system which is changing source of light and intensity of light on all parts of the ID card.

The second way is to test new methods of character segmentation with or even without our method which is depending on baseline detection, like using HMM in character segmentation or any other methods in order to achieve higher accuracies of character recognition.
The third way of development is to develop the system to be used for other purposes and be more generalized in the field of recognizing documents which have fixed locations of information like ID cards, license cards and passports, which will make the system suitable for a lot of governmental institutions and authorities.

The fourth way of development is to make the system deal with any printed documents with non-fixed information locations and different font sizes and types like business cards and OCR systems for books and papers, and the difficulty of this development is the existence of many font types which should be considered in the training phase, so the training phase will be very complex, then as an extra development, making android application that allows the businessman to take shots with his mobile camera for any business card, and the system extracts data and saves data in the database in the mobile and when the mobile connects with the personal computer of the businessman, he can export and import data to and from the database.
Finally the fifth way of development is to develop the system to deal with handwritten documents, and this development needs a huge training set to train the system to recognize handwritten sentences, and this system could be used with forms in any business company or governmental authority.

REFERENCES

[1] Rafael C. Gonzalez, Richard E. Woods and Steven L. Eddins, "*Intensity Transformation and Spatial Filtering*," in Digital Image Processing Using Matlab", 2009 by Gatesmark, LLC, pp.68–69.

[2] SaiCharan K, "*A Block DCT based Printed Character Recognition System*," http://www.cs.ucr.edu/~scharan/assets/Optical.Character.RecoRecogni.pdf, March, 2006.

[3] John G. Proakis and Dimitrics G. Manolakis, *Digital Signal Processing*. 3rd. Chapter 5, Prentice Hall, 2009.

[4] Edward Dougherty, "*Mathematical Morphology in image processing*", 2003, http://ebooks.spiedigitallibrary.org/book.aspx?bookid=159.

[5] JesalVasavada, and ShamikTiwari,"*A Hybrid Method for Detection of Edges in Grayscale Images*", I.J. Image, Graphics and Signal Processing, 2013, 9, 21-28.

[6] A. Zidouriand M. Sarfraz, S . A. Shahab and S . M. Jaf ,"*Adaptive Dissection based segmentation of printed Arabic text* ", Information Visualization, 2005, Ninth international conference on information visualization.

[7] M. Rashwan, M..Fakhr, M. Attia and M. EL-Mahallawy, "*Arabic OCR System Analogous to HMM-Based ASR systems* ", http://www.rdi-eg.com/Intro_to_NLP/Paper3.pdf 2007.

[8] AbdelwadoodMesleh, Ahmed Sharadqh, Jamil Al-Azzeh, MazenAbu-Zaher, Nawal Al-Zabin, TasneemJaber, AroobOdeh and Myssa'aHasn, "*An Optical Character Recognition*", Contemporary Engineering Sciences, Vol. 5, 2012, no. 11,Pages 521 – 529, http://www.m-hikari.com/ ces/ces2012/ces9-12-2012/meslehCES9-12-2012.pdf

[9] SaiCharan K., "*A Block DCT based Printed Character Recognition System*", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-2, Issue-6, May 2013, http://www.ijitee.org/attachments/File/v2i6/F0820052613.pdf.

[10] Mohammed Zekikhedhr,GeithAbandah,"*Arabic Character Recognition using Approximate Stroke Sequence*", (LREC 2002), http://gandalf.aksis.uib.no/ lrec2002/pdf/ws13/Khedher.pdf.

[11] Mansoor Al-A'ali and Jamil Ahmad, "*Optical Character Recognition System for Arabic Text Using Cursive Multi-Directional Approach*", Journal of Computer Science, Volume 3, Issue 7, 2007.

[12] A. Cheung, M. Bennamoun, N.W. Bergmann, "*An Arabic optical character recognition system using recognition-based segmentation*", Pattern Recognition, Volume 34, Issue 2, February 2001, Pages 215–233.

[13] H. Imtiaz and A.Fattah, "*A DCT-based Feature Extraction Algorithm for Palm-print Recognition*",Communication Control and Computing Technologies (ICCCCT), 2010 IEEE International Conference on information visualization, Pages 657 – 660.

التعرف على بطاقات الهوية المصرية باستخدام نظام خاص للتعرف الضوئى على حروف اللغة العربية

محسن رشوان- أميرة عبد الكريم شعبان- أشرف حسين عبد الرسول - اسراء شكرى عبد الفتاح -علا علاء الدين عبد النعيم
قسم الالكترونيات و الاتصالات الكهربية – كليه الهندسة - جامعة القاهرة

حسنين البرهمتوشى
استاذ بكلية الحاسبات و المعلومات– جامعة الملك عبد العزيز .

ملخص:
ان التعرف الضوئى على الحروف و خاصة حروف اللغة العربية لهو مجال بحثى صعب و ملىء بالتحديات و لكنه مهم للغاية لأنه يدخل فى العديد من التطبيقات المهمة التى تحتاج للتعرف على نص مستند ما اوتوماتيكيا من صورته .
هذا البحث يعرض مشروع نظام قارىء لبطاقات تحديد الهوية المصرية و الذى يقوم باستخراج المعلومات المهمة و الاساسية من صورة البطاقة الملتقطة و التعرف عليها ثم تحويلها الى قاعدة بيانات سهلة الادراج فى الحاسب الآلى .
هناك العديد من التطبيقات لهذا النظام فهو من الممكن ان يستخدم فى تسهيل اجراءات العملية الانتخابية ؛ حيث يكون من السهل التحقق من بطاقات المواطنين اتوماتيكيا و التأكد من الادلاء باصواتهم و بذلك يمنع حدوث تزوير او ما شابه .
ايضا من الممكن استخدام هذا النظام فى المنشآت الحكومية و المصانع التى تحتاج الى قواعد بيانات لعديد من المواطنين و التحقق من هويتهم. هذا البرنامج يستغرق تقريبا ١٦ ثانية للتعرف على بطاقة واحدة و هذا ما يوضح مدى واقعية تطبيق و استخدام هذا النظام فى الحياة اليومية و من المتوقع للنظام ان يؤدى بشكل افضل حين الانتقال من المرحلة الاكاديمية الى مرحلة تصنيع المنتج النهائى .

Towards Building Automatic Scoring System for Assessing Free-Constructed Responses

Nihal Al-Nazli[*1], Sameh AlAnsary[*2]

[*]*Nihal-Sameh Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University ElShatby, Alexandria, Egypt.*

[1]n_alnazli@hotmail.com

[2]Sameh.alansary@bibalex.org

***Abstract*— This paper provokes the area of the automated assessment of free constructed responses. The paper takes interest in investigating the different techniques and approaches used in automatic scoring systems thus trying to judge them according to their points of weakness and strength. It also discusses the most prominent automated marking systems based on these techniques and approaches. In addition, an ongoing research that seeks to design an automatic scoring system of Arabic short answers is highlighted.**

1 INTRODUCTION

Many academics consider the assessment process an important aspect of the learning process. Assessment through quizzes or exams is considered an essential part of the learning process and is considered by the examiners and teachers as an integral part of examining the learner's knowledge. A good assessment is the only way to test the learner's knowledge.

For this reason, Computer Assisted Assessment (CAA) systems were developed to support the assessment process through objective testing, such as multiple choice questions and fill-in-the-blank exercises. One of the first applications of computers in testing is the IBM model 805. It was used in the United States in 1935 for scoring objective test items. The idea was to reduce the costly procedures of scoring multiple-choice tests since the computerized scoring of the tests was thought to produce more reliable results in comparison to the previous hand-scored ones[1].

However, evaluating the learners' learning progress in this way is not enough to measure higher cognitive skills and limits the feedback that can be provided to the learners. Therefore, many academics insist on open- ended questions that require the learner to write free constructed responses. These types of questions require candidates to write one or more sentences or even one or more paragraphs.

These questions are highly regarded and integral part of the examinations and are also extensively used by teachers. A system that could partly or wholly automate valid marking of free text answers would therefore be valuable.

Advances in computational linguistics and the increasing penetration of computers in schools, in addition to the urgent need for including questions requiring a learners to state, suggest, describe or explain or elaborate their knowledge about a certain subject were the main reasons leading many educational institutions to start funding researches that helps including such questions in the automated examination thus leading the path to automated assessment of such questions.

Burstein, Wolff et al. (1999) state that there is a remarkable movement in the educational field to augment the conventional multiple-choice items with free-response items. The reason behind this movement is the large volume of tests administered yearly where hand-scoring of these tests with these types of items is costly and time-consuming for practical testing programs [2].

The large number of students and the educational need to make more exams puts such a heavy burden on academics to assess their learners (i.e. score their students answers). Therefore many institutes are currently working on natural language understanding systems which could be used for computer-assisted scoring of free constructed responses.

First experiments on automatic scoring consisted of question types of multiple choice, true or false with a clear absence of open –ended items from the examinations. Authors relate this absence to what they referred to as the "unsuitable" nature of open –ended question types for machine marking. In other words, marking using automated systems faces difficulty of coping with the myriad ways in which credit-worthy answers may be expressed [3].

Successful automatic marking of free text answers would seem to presuppose an advanced level of performance in automated natural language understanding. However, recent advances in natural language processing techniques have opened up the possibility of being able to automate the marking of free text responses typed into computer without having to create systems that fully understand the answers.

Research on the assessment of free constructed responses includes two areas of computer assisted assessment. One of them is the grading of essays, which is done mainly by checking the style, grammaticality, and coherence of the essay.

The other is the assessment of short student answers which are designed for short factual answers where the line between right and wrong is clear. In such a case, an automatic short answer grading system provides scoring based on content rather than style through comparing a student answer to one or more correct answers[4][5].
Valenti, S., F. Neri, et al. (2003), concludes the difficulties of grading essays in what was called the perceived subjectivity of the grading process. Many researchers claim that the subjective nature of essay assessment is perceived by students as a great source of unfairness due to the variation in grades awarded by different human assessors. Furthermore essay grading is a time consuming and quite costly activity as it requires one or more human judges. They suggest that this issue can be handled by adopting assessment systems that will automatically grade essays where such a system would at least be consistent in the way it scores essays. In addition to enormous cost and time savings that could be achieved when the system grades essays with a level of agreement in scoring to those scores awarded by human assessor [6].
Many academics suggest that the automatic scoring of open- ended questions with free-responses removes the burden of scoring (especially in large student numbers) in addition to removing bias and fatigue unfairness marking. Siddiqi and Harrison (2008) summarizes the benefits of fully automated scoring of free- responses by saying;

> "*Key benefits of automating free-text marking include time and cost savings, and the reduction in (ideally, the elimination of) errors and unfairness due to bias, fatigue (on the part of the human marker) or lack of consistency.*"[5]

The discussion elaborated in this section tried to cover the reasons behind using automatic scoring systems of free-constructed responses. The remainder of this paper will focus on answering the question of how to build these automatic scoring systems. As previously mentioned, the advance in the Natural Language Processing (NLP) technologies accelerated the development of automatic scoring systems for assessing free-text responses. In fact, there are more than 20 CAA systems relying on different techniques and applied to several domains. These different technologies and approaches are the main focus of this paper where the most prominent and successful of them will be discussed in details in section 2. Examples of the systems relying on those technologies will be discussed in Section 3. As for section 4, a light is shed on an ongoing research that aims at building a system for assessing Arabic short answers. The last section, section 5, will include the conclusion and points of further research.

2 APPROACHES AND TECHNOLOGIES

Advances in the computational linguistics field and the Natural Language Processing (NLP) technologies, in addition to the urgent need for including open-ended questions requiring learners to write free-constructed response led many educational institutions to support the use of automated assessment of free text answers.
Literature reports many successful attempts to reach fully automated scoring systems. Those systems differ in the techniques and the scoring approaches used. In this section, the most prominent techniques and approaches are discussed giving much attention on the advantages and drawbacks of using these approaches.
Many Linguists debate on the right classification of these techniques and approaches. Some made their distinction based on whether this technique or approach tries to classify the text or tries to understand the text. Others made their classification according to the method of evaluation whether it's based on style or based on content or both [7].
In this paper, classification of techniques and approaches is according to the level of NLP required. Therefore, two categories are distinguished namely shallow natural language processing and full natural language processing.

A. Shallow Natural Language Processing

This category is characterized by using statistical techniques for the lexical level. It includes all systems that rely on statistical analysis of one or more features of the text. They don't involve complex NLP techniques where the pre-processing of the text includes only a tokenization and part of speech tagging phase. All the systems using the shallow natural language processing involve a training phase in which the parameters of evaluation are identified [7].
The following are some of the techniques that can be considered as subcategories of the shallow natural language processing.

1) Keyword Analysis: It is a simple technique that looks for coincident keywords between the student's answer and the reference or model answer. One of the common models used is the Vector Space Model (VSM) and the N-gram analysis.

The VSM is originally used in text categorization and information retrieval where texts are represented as vectors in a hyperspace and dimensions correspond to words. A frequency with which a word appears in a certain document is replaced by weights, such as tf _ idf. The VSM compared documents by calculating the cosine of the angle of their

associated vectors. A lower cosine angle means a higher similarity with the reference answer and, therefore, a higher score. The VSM is used as an additional module in E-rater which will be discussed later.
The N-gram analysis is used to substitute the single keyword analysis, where sequences of N consecutive words are compared between the reference answer and the student's answer. This is the approach used in Willow system.
Research consider this approach to be limited and the reports refer to a drawback of using this approach as it fails in extracting a representation of the meaning of the student answer thus cannot deal with synonyms and polysemous terms[7].

2) Latent Semantic Analysis (LSA): It is a complex statistical technique that can be considered as an extension of the Vector Space Model discussed earlier. It is originally developed for indexing documents and information retrieval. However, it is then used in automated essay grading to measure similarity between the student and the reference answer by finding the hidden semantic relationships between words [7].

LSA has an underlying assumption that every document has an underlying semantic structure represented by its words, and this structure can be captured in a matrix.[8][9].
In the LSA based approach, a text is represented as a matrix with each row in the matrix representing a unique word and each column representing context. The frequency of the word is represented in each cell. The relevance of each word in the passage is then measured thus transforming frequencies into weights. This is done using tf _ idf or the x2 function where words that are equally common in every context are given low weights while high weights are given to words that are very representative of particular context[7][9].
The LSA then uses Singular Value Decomposition (SVD) method to reduce the dimensions or vectors of the matrix. Authors consider this reduction as the heart of the LSA approach because it allows the representation of the meaning and induces semantic similarities between words [10].
This reduction is critical due to its important role in representing the meaning of word through the context in which they occur. It is also crucial considering the number of dimensions it works on reducing where if the number is too small, too much of the information will be lost. Also, if the number is too big, limited or not enough dependencies will be drawn between vectors [9]. According to Lemaire and Dessus (2001),

"*A size of 100 to 300 gives the best results in the domain of language* " [10].

Research on the LSA technique describes it as a powerful method which seems to have high correlation with human graders. The reason behind this is that it can find semantic similarities between words even if they exist in different contexts and it can also find semantic similarity of documents even if they share no words. In addition, the semantic information is derived only from the co-occurrence of words in a large corpus of texts where it makes no use of human-constructed parsers, taggers, dictionaries, semantic networks, or other tools and also there is no need to code semantic knowledge by means of a semantic network or logic formulas[10][8].
Although LSA is quite robust, it cannot judge most matters of spelling and grammar as it measures only semantic similarity while syntax and morphology are completely ignored. However, some studies don't consider this issue as a drawback of the technique as they assume that human graders give much weight to content rather than style or mechanics [8].
However, a couple of disadvantages are reported. First is that the process of SVD is computationally expensive especially when decomposing a large matrix (e.g. a matrix generated from a corpus containing several million words) and can take hours or even days of computer time. Secondly, determining the number of dimensions by which to reduce the scaling matrix is such a difficult task. This is due the crucial factor considering the number of dimensions explained earlier as they mustn't be too little or too big [8].
The Apex Assessor is under pinned by LSA and will be discussed in details in section 3.

3) Analysis of surface linguistic feature: It is a statistically based approach that depends on detecting a list of surface features of the text and measures them. These surface features include both linguistic and non linguistic characteristics. The total number of words per essay (essay length), sentence length, and word length are examples of the nonlinguistic characteristics, while the total number of grammatical errors, the types of grammatical errors, or the kinds of grammatical constructions (e.g., passive or active) that appear in the text are examples of the linguistic characteristics [11].

After detecting the features, the text goes through a training phase that works on each feature in this list to discover their relative importance thus giving each feature the appropriate weight. Finally the scoring phase or the "Calibration phase" adjusts the weights to the optimal values [7].
Using the approach of surface feature analysis in automatic scoring system has an advantage of being cost effective. This is due to the fact that this procedure does not need any rubric or model answer to be prepared. Therefore, the preparation cost is minimal, and the scoring procedure itself is fairly rapid.
However, reports refer to some notable problems. First, such an approach of depending solely on surface features, having for example the length of the essay to be the most contributing factor of scoring, might be considered as a weak scoring criteria and might also cause unfairness marking specially when it comes to scoring short but to-the-point essays. Second, authors assume that this scoring criteria could be easily deceived or discovered by the learners as it is

very transparent and simple. They say that a student can deceive the system by writing a very long essay with non sense words and according to this approach he/she will receive a very high score.
Therefore, authors suggest that scoring systems based on surface characteristics could be used in conjunction with human scoring of essays as a "second rater". They further empathize that this approach is not used alone however it can be used in conjunction with other more complex manual or automated procedures or solving the problem through augmenting a more complex content-based analysis[11]. This is the case with the Project Essay Grader (PEG) that will be discussed in the following section and which is based partly on this analysis, along with using some additional NLP tools, such as a grammar checker, a POS tagger and a parser.

4) Text Categorization Technique: In this approach, the automatic scoring of free text answers is considered a classification task. In other words, instead of giving a numerical score, the classification is done using a discrete set of classes. Pérez-Marín, Pascual-Nieto et al. (2009) explains the procedure of this technique as follows

> "*The common practice is to have a set of predefined categories, such as good and bad, or a scale with N points indicating the degree of correctness. The purpose is to classify each student's answer in one of those categories.*" [7]

The Bayesian Essay Test Scoring sYstem (BETSY) is one of the automatic scoring systems based on the Text Categorization technique that will be discussed in section 3.

5) Information Extraction (IE): This approach depends on acquiring structured information from the free text. In other words, it identifies entities in the text and finds the relations between them. As the case for all shallow NLP techniques, the IE approach does not need a deep parsing of the text. This can be considered as an advantage of employing this approach in automatic scoring systems because it makes the approach easy to implement.

One of the common techniques used for the IE approach is the Pattern-matching technique. This technique is used to evaluate student's answers by breaking the text answer into concepts linked by relations binding these concepts represented in templates. These templates are then compared to their peer human-expert model templates to produce the final score [7]. The Automark, Automated Text Marker (ATM) and Schema Extract Analyse and Report (SEAR) are examples of the automatic scoring systems based on this approach. The Automark will be discussed as an example of using the IE approach in section 3.

6) Clustering: In this approach, the answers are used to form clusters. The clusters are formed by grouping the answers that share similar word patterns. The Intelligent Essay Marking System (IEMS), discussed in the following section, uses the clustering algorithm called Indextron to automatically assess the free-text students' answers [7].

B. Full Natural Language Processing

This category involves using complex natural language processing techniques that works on different linguistic level including lexical, syntax, semantics and/ or discourse processing levels. In this category several NLP tools are applied to achieve the goal of auto-marking of free-constructed responses [7].
A syntactic analyser is one of the NLP tools used to identify constituents of the input free-text and the syntactic dependencies between them. As for semantic analysis, it involves diagramming a sentence to a type of meaning representation to identify the role that constituents perform in the actions or states reported in the text (e.g. patient, agent, location and so on). Whereas discourse analysis focuses on how context impacts sentence interpretation and information extraction locates specific pieces of data from a natural language document [7][9].
Research shows that employing the full natural language processing techniques in assessing the student's answer is expected to improve the performance of the scoring systems and out performs the other shallow natural processing techniques. It also obtains a discourse and semantic analysis that helps in effectively assessing the student's answer.
However, using this deep and complex NLP techniques is hard to accomplish and the system performance is very dependent on the quality of the NLP tools. In addition, it is very difficult to port across languages [7].
One of the current systems underpinned by these techniques is the C-rater system (will be discussed in Section 3)
The MultiNet Working Bench (MRW) system uses a different approach that also falls under the full natural language processing category. It works by comparing semantic networks representing the answer of the student with the model semantic network of the human expert.

Concluding this section, we can finally say that there are several techniques used in order to automatically assess free-constructed students answers. Each technique or approach has its own points of strength and weakness. It depends on the goal of assessment and the available resources in order to choose the technique to be used. Several techniques can possibly be combined in order to take advantage of all of them and achieve a higher level of performance.

3 RELATED WORK

As mentioned earlier there are several systems that seek automatic scoring of free-constructed responses. Those systems vary according to the relying technique used in the scoring process. In the earlier section we discussed the different techniques used in the scoring process and examples were given one or more systems underpinned by those techniques. In this section, those systems are discussed in details.

A. E- Rater

The electronic essay rater (E-rater) is a software application designed by Jill Burstein and a team of researchers of the Educational Testing Service (ETS) to evaluate the quality of an essay[9][12]. Pérez-Marín, Pascual-Nieto et al. (2009) as follows;

> *"a writing analysis tool that automatically evaluates, and scores essays written in English* "[7]

E-Rater was developed in the mid 1990's to produce holistic scores evaluating essay based on features of effective writing such as organization, sentence structure and content [8]. In the holistic scoring approach, scores are given based on the total impression of the essay taking into account all aspects of writing as specified in the scoring guide [13].

E-rater uses NLP tools to extract linguistic features of discourse structure, syntactic structure and vocabulary usage (domain analysis). These features are used to identify specific lexical & syntactical cues that are used in analysing the data [6][9].

E-rater adopts a corpus based approach in model building thus training its engine with a set of human graded essays based on unedited text corpora representing the specific genre of first-draft essay writing [6][13].

E-rater searches the essay to be graded for features reflecting the writing skills of the student. Therefore, E-rater architecture design consists of three independent modules for feature recognition. In other words, these modules identify features that may be used as a scoring guide criteria for syntactic variety, the organization of ideas and the vocabulary usage through the essay. These modules are syntactic module, discourse module, topic analysis module. These modules provide outputs for additional two modules namely model building and scoring module [13].

In the Syntactic module, the text is processed using a part of speech tagger and syntactic chunker [13]. Then E-rater uses NLP tools to capture syntactic variation in the essay. Dikli (2006) describes those variations as follows

> *"In order to capture syntactic variety in an essay, "a parser identifies syntactic structures, such as subjunctive auxiliary verbs and a variety of clausal structures, such as complement, infinitive, and subordinate clauses*"[9]

The discourse module then uses a conceptual framework of conjunctive relations including cue words, terms and the output syntactic structures from the previous modules to identify discourse-based relations and organization of essay [9].

In the Topic analysis module, the essay is evaluated for vocabulary usage and topical content using vocabulary content Analysis and Vector Space Model(VSM) technique explained in Section 2.This technique assesses the essay's content in terms of similarity with pre-scored essays[9][12].

The Graduate Management Admission Test (GMAT) is one of the computer based delivery tests that adapts the holistic essay scoring approach. Therefore, Educational Testing Service (ETS) was encouraged to implement e-rater for operational scoring of the GMAT Analytical writing Assessment (GMAT AWA) in 1999. Since then, over 750,000 GMAT essays have been scored, with e-rater and reader agreement rates consistently above 97% [13].

B. Apex Assessor

Apex assessor is an interactive learning environment created in the year 2000 by Dessus, Lemaire and Vernier in the Laboratorie des Sciences de L'E´ducation in the Universite´ Pierre-Mende´s in France[7][8].

It is a web based application originally developed to assess essay written in French language using Latent Semantic Analysis technique discussed earlier. The assessment is based on content rather than style, therefore LSA was implemented to compare the essay to be graded to the text of a given topic on semantic basis.

According to Apex's authors, the student is engaged in an iterative improvement process where the student has the chance of re-writing his/her essay more than once after receiving feedback. Lemaire and Dessus (2001) elaborates this idea as follows;

> *"The environment is designed so that the student can select a topic, write an essay on that topic, get various assessments, then rewrite the text, submit it again, etc."* [10]

The Apex assessor uses the LSA technique to provide three types of assessment on the student's essay namely, content based assessment, outline assessment and coherence assessment. The system compares the students LSA representation to the LSA representation of the text on the topic thus measuring the semantic similarity between the two representations and identifying how well the notions of the student's text answer is semantically similar to each notion of the selected topic. A student has the chance to modify his/her text after reading any of these three evaluation texts and resubmits his text[10].

To system was tested against 31 essays of a graduate course on sociology of education. Comparing the results to the teachers' grade, the Apex results showed 59% correlation with p,0.001 [7]. Other reports on the Apex's performance state that it has good correlation with human scores for content (r = 0.64) and overall essay quality (r = 0.59) [8].
Reports recommend this approach in a distance learning context since it is a web based application that enables students to connect to the system and freely submit essays. In addition, the system evaluates the essays as many times as required by the students without the need for the teacher to code any domain knowledge. On the other hand, Apex evaluation of content has some reported limitations. One of them concerning very short student texts (i.e. cases where the student writes just a few words) that some graders might want to score it with very low scores. However, some reports states that the Apex content-based assessment could yield a high score. Another problem concerning the core of the LSA technique underpinned in the system. As explained earlier, LSA neglects the syntax of the text therefore the Apex system has no way of detecting when a sentence has syntactic errors or when some common words are missing [10].

C. Project Essay Grade

Ellis Page developed Project Essay Grade (PEG) in the year between 1965 and 1966 in the University of Duke in USA. It is reported in literature as the first serious attempt at scoring essays by computer [7][8].
Page claims that his system does not intend to "understand" the content of the responses. He and other assisted developers of the system further argue that understanding the content is very difficult, impractical, and moreover it is considered an unnecessary goal, especially with large-scale assessments [12]. Therefore, he used a statistical approach of surface linguistic analysis explained earlier that focuses on the style of the essay. Thus, an essay is graded on the basis of writing quality, taking no account of content [6].
The approach used for PEG is based on the concept of "proxes" and "trins" in generating the score of an essay. The term "trins" is used to refer to the intrinsic variables such as fluency (such as: counts of prepositions, relative pronouns and other parts of speech), diction (such as: variation in word length), grammar, punctuation, etc. On the other hand, "*Proxes"* represents the number of words in the essay (essay length) [6][9].
The system contains two stages; a training stage and a scoring stage. In the training stage, Proxes are calculated from a set of training essays. In the scoring stage, proxy variables are determined for each essay and then transformed and used besides the given human grades for the training essay in a standard multiple regression to calculate the regression coefficients (weights) [6][9]. Page introduced 28 different proxes such as the title, the average sentence length, the number of paragraphs, the punctuation and the number of prepositions in the first version of the PEG system in 1966[7].
Reports of later implementation of the PEG system state that in 1990 it used grammar parser and a POS tagger to improve the proxes discovery. A later enhancement of the system in 2002 reports that the system currently includes content, organization, style, mechanics and creativity assessment.
Research on PEG suggests that it is suitable for most types of essays, achieving 87% correlation between its scores and human ones [7]. However, in addition to the drawbacks of the surface analysis technique discussed in section 2, another criticism to the PEG system is that it needs to be trained for each essay set used & it further needs a relatively large training data ranges from 100 to 400 sample essays [9].

D. Bayesian Essay Test Scoring sYstem

Rundner and Liang developed the Bayesian Essay Test Scoring sYstem (BETSY) at the college Park of the university of Maryland. The development phase was between the year 2001 and 2003 with the funds from the U.S. department of Education [7].
The system is designed to classify text based on trained material. According to its authors, the system's goal is to classify an essay using a four point nominal scale (e.g. extensive, essential, partial, unsatisfactory). This classification is done using text categorization techniques that utilizes a large set of features taking into account both content and style[6].
In order to learn how to classify new documents, BETSY goes through a number of steps in the training phase. It train words, eliminate uncommon words, determine stop words, evaluate database statistics, train word pairs. It sometimes score the training set and trim misclassified training texts. After the training, BETSY can be applied to a set of trial texts to determine classification accuracy [9].

The system was used to assess Biology items for the Maryland High School and it achieved about 75- 80 % accuracy. Furthermore, Rudner and Liang say that their system could be applied to any text classification task [7]. Despite the large training data needed to train the BETSY, it is claimed that it includes the best features of PEG and e-rater along with its own essential characteristics. It is an open source system that is simple to implement and easy to explain to non-statisticians. In addition, the system proved to be effective dealing with short essays and various content areas [9].

E. Automark

Automark is a software system that aims at automatic computerized marking of free text answer to open-ended questions. The system has been under development for almost three years. It was first created as an academic work by Mitchell, Russell, Broomhead and Aldridge from the University of Liverpool and Brunel University in UK in 1999. The development continued for 3 years till 2002 when they founded their own company namely Intelligent Assessment Technologies. It was only then when the system was employed in a commercial e-learning product called "Exam online" and was available for registered users only[6][7].
The system was designed to assess student essays based on both the style and content of the essay thus indicating whether it's acceptable or not according to criteria specified by the teacher [7].
Automark adopts IE techniques in addition to some NLP techniques to accomplish its assessment goal. The system aims at performing an intelligent search in the student's free text response according to predefined computerized mark scheme templates. These mark scheme templates are represented in the form of syntactic-semantic templates. This way of performance resembles the way of humans when marking free-text responses. In order to resembling human markers, the system incorporates a number of processing modules that attempts to identify the student's understanding expressed in his/her free-text response thus ignoring some mistakes. These mistakes include misspelling, typing errors, some syntax and semantic mistakes [14].

The Automark system was used at a number of higher education establishments. One of those was an online java test for first year engineering student at the Brunel University. It has also been tested on National Curriculum Assessment of Science for eleven years old pupils. Valenti, Neri et al. (2003) describes the testing environment and the performance of the Automark as follows

> *"Automark has been tested on National Curriculum Assessment of Science for eleven years old pupils. The form of response was: single word generation, single value generation, generation of a short explanatory sentence, description of a pattern in data. The correlation achieved ranged between 93% and 96%."*[6]

F. Intelligent essay marking system(IEMS)

The IEMS is a an automatic essay grading system developed in 2000 by Ming, Mikhailov and Kuan at the NGEE ANN Polytechnic in Singapore. The system is based on clustering algorithm called Indextron which performs pattern matching based on Pattern Indexing Neural Network [7]. The system aims on providing both summative and formative assessment therefore, it can be used both as an assessment tools and for diagnostic and tutoring purposes in many content-based subjects. The essay grading is based on qualitative type of questions rather than numerical type[6].

Research recommend that IEMS is embedded in an intelligent tutoring system as it grades answers rapidly and provides students with immediate feedback quickly. In addition, the feedback given to students tends to be precise whereby students can learn where and why they had done well or not made the grade [6].
The system was tested as an experiment of evaluating students' answers summarizes an 800-word passage entitled ‘Crime in Cyberspace’ in the fulfillment of the Project Report Writing module. The experiment involved 85 students of third-year Mechanical Engineering. They were asked to write a summary of not more than 180 words about the text. The results of the IEMS were promising as it obtained a correlation of 0.8 with the teacher’s scores [6][7].

G. C-rater

C-rater is an automatic scoring system developed by the American Educational Testing Service (ETS) organization. It is used to assess short free-text student answers related to content based questions such as those such as those that may appear in a textbook’s chapter review section. These types of question requires answers that range in length from a few words to approximately 100 words [15].
The system's goal is to automatically score short student answer thus measuring the student's understanding of certain concepts without considering his/her writing skills (i.e. it focuses on meaning thus tolerating form errors). Valenti, Neri et al. (2003) explains this by saying

> *"C-rater is aimed to score a response as being either correct or incorrect. This goal is achieved by evaluating whether a response contains information related to specific domain concepts; if the response expresses these concepts it is rated as correct, otherwise it is rated as incorrect without any regard to writing skills." [6]*

To achieve this goal, C-rater exposes the text answer to deep natural language processing using a set of NLP tools. These tools extract the linguistic features from both the model answer and student answer. The C-rater recognizer task here is to recognize a correct response with all it's possible variation forms. Whether these variations are syntactic or semantic (i.e. using synonyms or similar terms or the variations are due to misspelling or different inflections of a word or due to the use of pronouns in place of nouns).
Therefore, C-rater engine applies a sequence of natural language processing steps on the text answer. These steps include:

- Correcting spelling mistakes and typing errors.

- Determining grammatical structure of each sentence.
- Resolving pronoun reference and analyzing paraphrases.[16]

C-rater is very reputational for high scoring accuracy of short answers. Authors relate this high level of accuracy to its underlying full natural language processing approach described earlier in section 2. C-rater was used in formative low stakes tests on both small scale and large scale. In both cases C-rater managed to achieve a high level of agreement with human raters. Reports show that C-rater achieved over 80% agreement with the instructor when utilizing the system in a small-scale study with a university virtual learning program. It was also used in a large-scale assessment to score 170,000 short-answer responses to 19 reading comprehension and five algebra questions, the result was 85% accuracy [7].
Despite this high level of scoring accuracy and agreement with human raters, C-rater has some drawbacks. One of the most prominent drawbacks is the need for a large number of scored responses in order to build the system's model. Reports states that this number might exceed a 100 scored responses where in some cases as 200 was found to be insufficient. In addition, the fact that C-rater scores questions according to a set of finite concepts is might be considered a limitation. This is due to the extensive manual effort required to build models and therefore, experiments are done to introduce interactive machine learning techniques in the process of Model building. Another limitation is that C-rater fails to score open-ended questions asking for opinions or a student's own experience.[1]

4 ARABIC AUTOMATIC SCORING SYSTEM

This section sheds some light on an ongoing research that aims at building a system for automatically assessing student's short responses to academic questions in Arabic language. It is considered a pioneer study since most of the work on automatic assessment for free-constructed responses is designed for English language, with few studies on other languages not including Arabic.
The significance of the study is that working on Arabic will enrich Arabic Natural Language Processing (NLP) resources and tools which will positively affect Arabic NLP. In addition, using Arabic automatic scoring system will benefit Arabic E-learning systems.
The study aims at developing a scoring system that assesses Arabic short answers to academic questions trying to overcome difficulties faced by other grading systems. The system built uses questions that require candidates to write one or two sentences at the most. Therefore, the primary step to collect the study data is to construct a kind of exam to which participants can respond, and the pool of their responses is used to construct the corpus of data required for the purpose of this research.
The rationale used in the construction of the study tool has observed the following criteria:

- The questions are professionally designed and pretested by qualified academic staff.
- The question difficulty level is within the reach and ability of the average freshman student.
- The Arabic form of the test question is revised and edited after the process of back translation.
- The test form is further reviewed by linguists from the teaching staff where the content, especially professional terminology, are focused on.

To design the test questions used for data collection, past years questions used mainly for first year student in General Linguistics, Phonology, and Anatomy are reviewed and compiled. We also went through some of the questions used in the text books used in the students' syllabus. The chosen questions are all translated into Arabic language and are reviewed and approved by two faculty members who teach General Linguistics and Anatomy.
On the bases of professional discussion and approval of the two specialists from the staff members, 30 academic questions are selected for the purpose of this study. These questions are seen by the researcher to meet the requirements of the above explained rationale and satisfy all the criteria required in the construction of the research tool.
Using the ASP.NET programming language through the Visual Studio application software building and building the data on SQL Server database engine, the 30 questions of the constructed exam with exam directions are programmed which helped research participants directly input their responses into the computer. An example of the questions used in the constructed exam with the interface used to submit the answers is shown in Fig. 1

[1] Leacock, C. (2004). "Scoring Free-Responses Automatically: A Case Study of a Large-Scale Assessment" http://www.ets.org/Media/Research/erater_examens_leacock.pdf

Figure 1: The interface of the proposed system with question examples

The research participants are all enrolled to study at the Department of Phonetics, the Faculty of Arts of Alexandria University. They are enrolled at different levels of study, from freshman to senior students.

The constructed exam is administered in groups during the official final exam period, for 45 minutes. The exam is administered under the supervision of the researcher in the presence of two staff members who helped answer the students' queries. The Model Answer and a Cut- Off point scale is provided by the academic staff members who teach the participants in their major subjects.

Both the student's answers and the model answers undergo a pre-processing phase using the Universal Networking Language (UNL). The UNL is an artificial language developed to enable computers to represent and process information across language barriers [17][18]. In other words, it is designed to replicate the functions of natural languages. The goal was to use UNL to describe all information and knowledge conveyed by natural languages for computers.

The UNL was particularly chosen for the purpose of this study as it expresses information and knowledge in the form of semantic network. The UNDL Foundation that is responsible for the development and management of the UNL Program describes the UNL expression as follows;

> *"The semantic network of the UNL is a directed graph. Its nodes are UWs or hyper-nodes (or 'scope' as it is commonly called) representing concepts. Its edges are Relations between concepts. Concepts can be annotated by Attributes. Such a semantic network of the UNL is called a 'UNL Expression or 'UNL Graph' "* [18]

This way of semantic representation makes the natural language represent non-ambiguous and have no redundant expressions. In addition, UNL is fully compositional where the UNL expressions must derive their semantic value thoroughly from their components that are explicitly defined in the UNL Knowledge Base. Accordingly, figure of speech, such as metaphor and metonymy are not allowed in the UNL. Instead they are represented, in UNL, by their intended meaning [17]. Those properties of UNL in addition to others, such as being declarative and complete was thought to be very useful and efficient for the purpose of the research, where all the data will be UNLized and represented through semantic network, or UNL graph. UNLization is the process of representing the content of natural language and providing an infrastructure of information and knowledge into the structure of UNL. This UNL structure (UNL graph) is made of three different types of discrete semantic entities which are the Universal Words (UW) that are the nodes in the semantic network. In addition to Universal Relations which are the arcs linking UW's and finally Universal Attributes that are used to instantiate UW's [17].

For the purpose of this study and in order to make the data UNLized, the Interactive ANalysis (IAN) system is utilized. IAN is a web application developed in Java and works as a natural language analysis system that can represent natural language sentences as semantic networks in the UNL format. It's currently available at the UNLdev[17]. All the data to be processed conformed to the IAN's requirement including UNL analysis dictionary, UNL analysis transformation

grammar, UNL analysis disambiguation grammar which are all consistent with the UNL specs. Figure 2 represents an example of a student answer UNLized using IAN and the output is shown both on the IAN interface and the UNL editor.

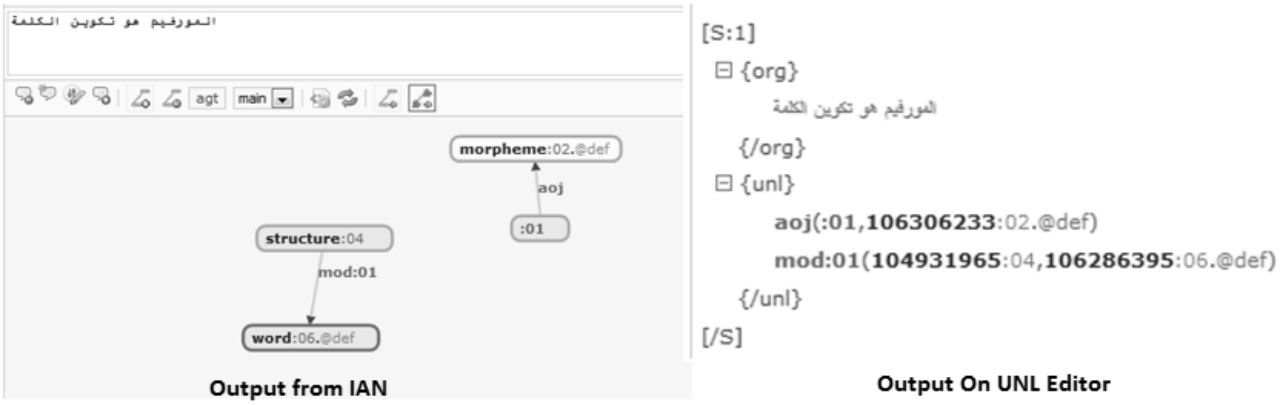


Figure 2: A student answer in Arabic UNLized using IAN

The matching and scoring algorithm is under construction, and the whole system performance will be evaluated in comparison to its pear scoring systems mentioned earlier.

5 CONCLUSIONS

Automated scoring of free-constructed responses is increasingly in demand since it eliminates the marking burden on staff and even eliminating marking errors and unfairness due to bias, fatigue or lack of consistency on the part of the examiners. Systems of such kind proved to be helpful and saves time and effort.

At the beginning, the thought of including open-ended question types in such system was impossible and impractical. However, advances in the computational linguistics field and the NLP tools made the automatic scoring of such question types possible. Several CAA systems were reported in literature, each utilizing a different techniques and marking approach. The goal beyond the assessment system is the motive toward choosing the suitable marking algorithm.

Although many CAA systems are discussed, most of them work on English language with a complete absence of such systems on Arabic. Therefore, the system under study is expected to enrich the Arabic NLP by adding a new system developed specially for Arabic language.

REFERENCES

[1] Parhizgar, S. (2012). "*Testing and Technology: Past, Present and Future*." Theory and Practice in Language Studies 2(1): 174-178.

[2] Burstein, J., S. Wolff, et al. (1999). *Using lexical semantic techniques to classify free-responses*, Springer.

[3] Sukkarieh, J. Z., S. G. Pulman, et al. (2003). *Auto-marking: using computational linguistics to score short, free text responses*. Paper presented to the ٢٩th annual conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

[4] Mohler, M. and R. Mihalcea (2009). *Text-to-text semantic similarity for automatic short answer grading*. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics.

[5] Siddiqi, R. and C. J. Harrison (2008). *On the automated assessment of short free-text responses.* Paper presented at the 34th International Association for Educational Assessment (IAEA) Annual Conference, Cambridge, UK.

[6] Valenti, S., F. Neri, et al. (2003). "*An overview of current research on automated essay grading"*. Journal of Information Technology Education 2: 319-330.

[7] Pérez-Marín, D., I. Pascual-Nieto, et al. (2009). *"Computer-assisted assessment of free-text answers."* The Knowledge Engineering Review **24**(04): 353-374.

[8] Miller, T. (2003). *"Essay assessment with latent semantic analysis."* Journal of Educational Computing Research **29**(4): 495-512.

[9] Dikli, S. (2006). *"An overview of automated scoring of essays."* The Journal of Technology, Learning and Assessment **5**(1).

[10] Lemaire, B. and P. Dessus (2001). *"A system to assess the semantic content of student essays."* Journal of Educational Computing Research **24**(3): 305-320.

[11] Kaplan, R. M., S. Wolff, et al. (1998). "*Scoring essays automatically using surface features."*. Educational Testing Service Research Report, GRE Board Professional Report No.94-21P, 98-39

[12] Yang, Y., C. W. Buckendahl, et al. (2002). *"A review of strategies for validating computer-automated scoring."* Applied Measurement in Education **15**(4): 391-412.
[13] Burstein, J., C. Leacock, et al. (2001). *"Automated evaluation of essays and short answers*." In Proceedings of Fifth Annual International Computer Assisted Assessment Conference Loughborough University, UK, Learning & Teaching Development, Loughborough University, 41-45.
[14] Mitchell, T., T. Russell, et al. (2002). *"Towards robust computerised marking of free-text responses."* In M. Danson (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughboroug University, Loughborouh, UK.
[15] Sukkarieh, J. Z. and S. Stoyanchev (2009). *Automating Model Building in c-rater*. Proceedings of the 2009 Workshop on Applied Textual Inference, Association for Computational Linguistics.
[16] Gomaa, W. H. and A. A. Fahmy (2012). *"Short Answer Grading Using String Similarity And Corpus-Based Similarity."* In International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 3, No. 11, 2012.
[17] http://www.unlweb.net/wiki/Introduction_to_UNL
[18]http://www.undl.org/index.php?option=com_content&view=article&id=46&Itemid=58&lang=en

نحو بناء نظام تصحيح آلي لتقييم الإجابات الحرة الغير مقيدة

نهال النظلي[١]، سامح الأنصاري[٢]

قسم الصوتيات واللسانيات، كلية الآداب، جامعة الاسكندرية
الشاطبى، الاسكندرية، مصر
[1]n_alnazli@hotmail.com
[2]Sameh.alansary@bibalex.org

ملخص:
تهتم هذه الورقة البحثية بموضوع التقييم الآلي للإجابات الحرة الغير مقيدة، حيث تبحث التقنيات والأساليب المختلفة المستخدمة في أنظمة التقييم الآلية. ولما كان الهدف هنا هو تقييم هذه الأنظمة، فإن الدراسة اهتمت بفحص نقاط القوة والضعف لكلٍ منها.
وعلاوة على ذلك فقد ناقشت هذه الورقة البحثية أبرز انظمة التصحيح الآلي والتي تقوم بشكل أساسي على هذه التقنيات والمناهج . وأخيراً فقد سلطت الدراسة الضوء على أحد الأبحاث القائمة التي تعمل على تصميم نظام تقييم آلي للإجابات القصيرة المكتوبة باللغة العربية.

Towards Building a Rule-Based Dependency Parser for Modern Standard Arabic

Rehab Arafat[1], Sameh Alansary[2]

Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University
ElShatby, Alexandria, Egypt.

[1]rehab.arafat@gmail.com

[2]sameh.alansary@bibalex.org

***Abstract*—In this paper the researcher is offering the on-going steps to build a dependency parser for Modern Standard Arabic sentences. Dependency grammar is offering a linguistic framework for free word-order languages such as Arabic. The main idea of Dependency parsing is the structure of a sentence is consisting of lexical items attached to each other by binary asymmetrical relations, known as dependency relations. The origins of dependency grammar can be traced back to the old Arabic literature and some of the Latin and Greek grammarians. This paper will discuss some of the basic issues of dependency grammar and the on-going steps to parse an Arabic corpus with the Interactive ANalyser of the ULN framework (IAN) according to the principles of Dependency grammar**

1. INTRODUCTION

Arabic language came in the sixth rank in the world's league table of languages, with an estimated 186 million native speakers (2009). The modern form of Arabic - Modern Standard Arabic (MSA) - is a simplified form of classical Arabic, and follows the same grammar. The main differences between classical and MSA are that MSA has more modern vocabulary, and does not use some of the complicated ones.

The process of parsing would be the backbone for other applications in the field of natural language processing, such as machine translation, text mining, information retrieval, texts summarization, speech processing, and others. The term "parsing" is used to refer to the process of building automatically syntactic analysis of sentences according to the grammar of a given language[1].

Parsing Arabic sentences faces many challenges. These challenges originate from the omission of diacritics in the modern written form of Arabic. Another challenge is the free word-order nature of the Arabic sentences. The presence of an elliptic personal pronoun also represents another challenge. The morphological richness of Arabic words and the inflectional nature of Arabic also form a real challenge.

An efficient syntactic model should be adequate in two areas; in description and in implementation. In other words, it should be described linguistically in order to be applied computationally. So, the aim of the research is to discuss the linguistic framework of Dependency Grammar, the history of its roots, the different theories of Dependency grammar and the pros and cons of applying Dependency grammar generally and in Arabic specially. After this survey it would be possible to suggest a convenient plan in order to build a parser system for Modern Standard Arabic based on Dependency Grammar. This would be achieved through the following stages:

- Make a broad survey for the different techniques and methodologies of parsing.
- Defining why to implement Dependency Grammar in rather than Constituency.
- Exploring the previous efforts in building Dependency parsers for Arabic, and how these parsers were evaluated quantitatively and qualitatively.
- Building a dependency parsing rules for Modern standard Arabic sentences.
- The parsers will include also an Arabic morphological Rules and an adequate dictionary.
- In addition to building the Dependency Grammar rules, a disambiguation module will be built to enhance the output of the parser.

- Presenting an accurate methodology of evaluating the output of the parser.

The flow of the work will include three basic modules as in Fig. 1:

- An Arabic morphological analyzer
- The lexicon structure
- The grammar rules.

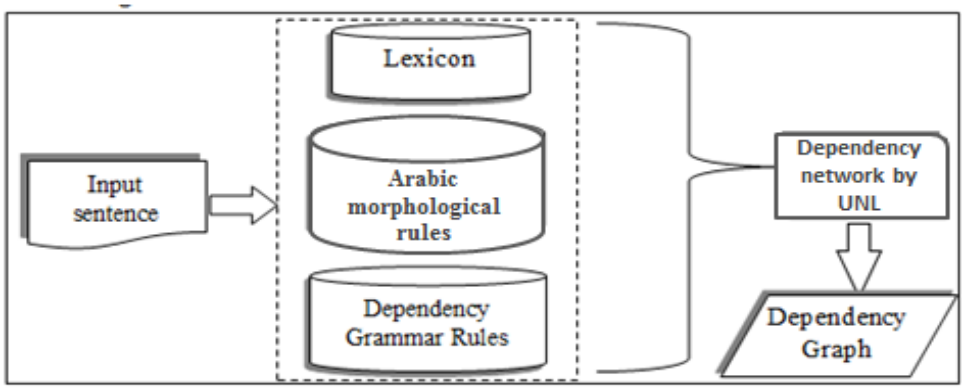


Figure 1: Basic components of the parsing system

2. THE DEBATE OF SYNTAX AND SEMANTICS AUTONOMY

Syntax is playing a critical role in defining the systematic relationship between the syntactic structure of the utterance and its meaning; you must know which words modify which other words in order to know the correct interpretation of the sentence. Chomsky and his students adapted the separation between syntactic and semantics. He claimed that syntax and semantics are essentially two separate fields and they should be treated independently "*I think that we are forced to conclude that grammar is autonomous and independent of meaning*" [2]. Chomsky viewed that the syntactic rules are formed with no reference to the meaning. As the generative approach assumed, the two disciplines must be kept separate; firstly it is a must to explain the structure of a sentence and the rules that governed that structure, and then define the meaning of the sentence. However, there is a significant connection between the structure of the sentence and it is meaning, as

Chomsky noted that "*It is reasonable to suppose that the needs of communication influenced [language] structure*"[3]. Our aim here is to adapt the dependency grammar approach as it attempts to connect the lexical items of the sentence in a way that would simulate the human brain.

3. WHAT IS DEPENDENCY GRAMMAR

There are four types of dependencies; semantic, syntactic, morphological and prosodic dependency. It is very important to distinguish clearly between types of dependencies, as Semantic dependencies are usually overlapping with syntactic dependencies. As Nivre (2009) mentioned [4], Dependency Grammar is mainly assuming that the structure of a sentence is consisting of lexical items, these lexical items are attached to each other by binary asymmetrical relations, and these relations are known as dependency relations or dependencies. The dependency relation is established between two nodes; the head and the dependent.[i] There are set of criteria for controlling the establishment of dependency relations between lexical items, and for defining the head and the dependent in these relations.

A dependency relation holds between a head and a dependent.

[i] Alternative terms: *governor* and *modifier*, *parent* and *child*.

Alternative terms in the literature are governor and regent for head and modifier for dependent. The head of a sentence is usually taken to be the tensed verb, and every other word is either dependent on the sentence head, or connects to it through a path of dependencies.
A dependency representation is a labeled directed graph, where the nodes are the lexical items and the labeled arcs represent dependency relations from heads to dependents. Fig. 2 illustrates a dependency graph, where arrows point from heads to their dependents for the sentence "Economic news had little effect on financial markets".

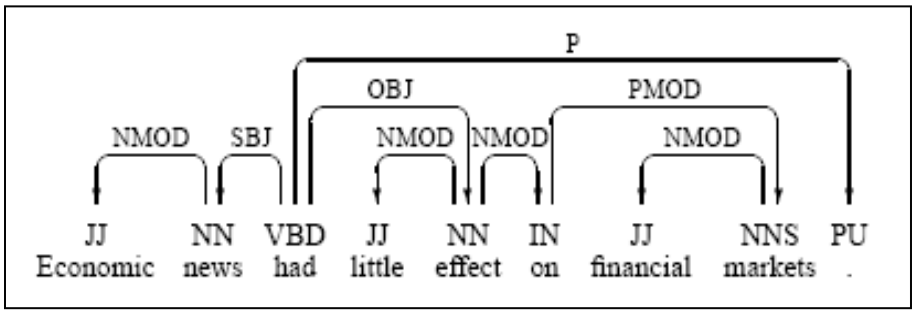


Figure 2: Dependency structure for English sentence from the Penn Treebank.

In Fig. 2, arrows point from heads to their dependents; labels indicate the grammatical function of the dependent as subject, object or modifier.
But there are basic criteria for establishing dependency relations, and for distinguishing the head and the dependent in such relations. Such criteria have been discussed not only in the dependency grammar tradition, but also in other frameworks where the concept of the syntactic head plays an important role, including all constituency-based frameworks that subscribe to some version of X bar theory [5].

4. ROOTS OF DEPENDENCY GRAMMAR

Although Phrase Structure based theories have been discussed and applied widely, but recently there is a great tendency towards using Dependency grammar as a linguistic framework and as a parsable technique. And many linguists lately considered Dependency Grammar to be inferior to phrase structure theories. However the roots of modern Dependency grammar are dating back to the French linguist Lucien Tesnière, the concepts of Dependency grammar have already been discussed by traditional grammarians as Covington argued. Some concepts of dependency grammar have been discussed by traditional grammarians in the Arabic literature. It was discussed also in the Indian Grammar by Pānini and in the Latin and Greek grammar.

5. DEPENDENCY VERSUS CONSTITUENCY

There are many arguments that recommend using dependency rather than constituency in syntax and parsing for its capacity of its representation.
Dependency approach differs from Constituency in the way in which the structure is forming. According to constituency, a group of words are forming a larger unit which constitutes the "constituent" or the "phrase". However in Dependency Grammar, there is a dependency relations between two words and there is no other units larger than units. The dependency graph is a hierarchal links between the elements of the sentences. The hierarchal structure is c1onsisting mainly of a set of binary relations. However, the Phrase structure tree is a grouping of chunks that constitute a lager unit. The word order in the PS tree is linear while the dependency graph is two dimensional. The dependency

graph is consisting of terminal nodes linked by a set of dependency relations, while the PS tree contains both terminal and non-terminal nodes.
The sentence "حجز علي التذكرة من مكتب الطيران" will be represented as in Fig. 3 according to dependency grammar.

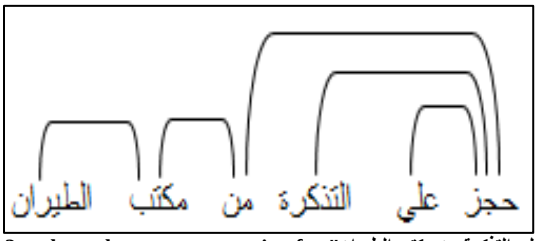


Figure 3: a dependency representation for "حجز علي التذكرة من مكتب الطيران"

In previous diagram the verb "حجز" is called the head, and the word "علي" is the dependent of that head. The head "حجز" had more than one dependent, and this allowed. But it is not allowed for the dependent to have more than one head.

Fig. 4 illustrates the phrase structure representation of the same sentence. As shown in that Fig. 4the sentence (s) is consisting of some constituents or phrases (VP, NP, NP and PP).

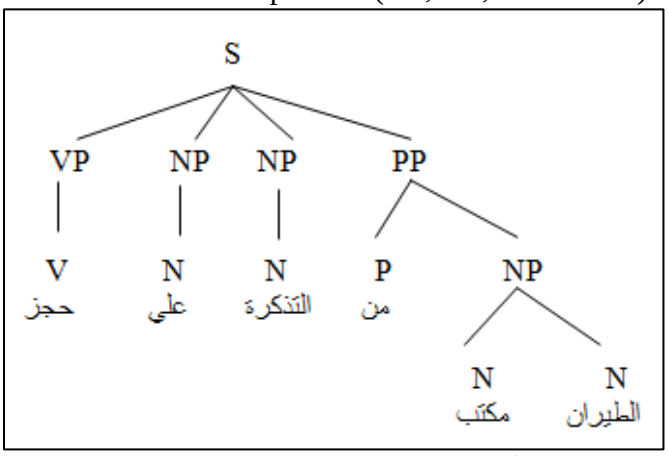


Figure 4: PSG representation for "حجز علي التذكرة من مكتب الطيران"

We could admit that both DG and Constituency are dealing with the term "head". However the two approaches are dealing with the notion of the "head" in different ways. According to the constituency approach, the definition of the "head" refers to the head of the phrase for each grammatical category as the following:

- Noun Phrase (NP): Noun (N) is the head.
- Adjective Phrase (AP): Adjective (A) is the head.
- Verb Phrase (VP): Verb (V) is the head.
- Prepositional Phrase (PP): Preposition (P) is the head.

On the other hand, usually, the head is the verb of the sentences according to dependency grammar.

6. DEPENDENCY PARSING

Dependency parsing is the process of the automatic syntactic analysis of natural language sentences by means of the theoretical linguistic framework of dependency grammar [3]. Recently, dependency parsing has attracted considerable interest from researchers and developers in the field for many reasons. One of these reasons is that dependency-based syntactic representations seem to be more computationally applicable more that Phrase Structure Grammar, in other words, it would be more useful in many applications of language technology, such as machine translation and information extraction because of their transparent encoding of predicate-argument structure. Another important point regarding DG is that it is better suited than phrase structure grammar for languages with free or flexible word order; it makes it possible to analyze typologically diverse languages within a common framework.

So far, some theoretical tradition of dependency grammar was reviewed. Now the light will be shed on the main topic of this research, the computational implementation of syntactic analysis based on dependency representations, i.e. representations involving lexical nodes, connected by dependency arcs, and labeled with dependency types. The connections between theoretical frameworks and computational systems are often rather indirect for dependency-based analysis [6]. In discussing dependency-based systems for syntactic parsing, Carroll [7] distinguished two broad types of strategy, the grammar-driven approach and the data-driven approach, and these approaches are not mutually exclusive.

7. MAIN ISSUES IN THE CONCEPT OF DEPENDENCY

The basic assumption of Dependency Grammar is that the structure of the sentence is consisting of lexical items, these lexical items are attached to each other by binary relations, and these relations are known as dependencies. The common formal property of dependency structures, as compared to representations based on constituency is the lack of phrasal nodes, and the syntactic structure of sentences depends on binary asymmetrical relations holding between lexical elements. However, there are also important differences and issues to be discussed here.

One important issue is the Mono-stratal versus multi-stratal frameworks. Some dependency theories are distinguished by its notion of the layers (Multi-stratal). In other words, some theories depend on only one syntactic representation (Mono-stratal), while other theories depend on many layers of representations (Multi-stratal). Most of dependency theories are multi-stratal. This section will discuss the difference between the theoretical frameworks of dependency grammar that rely on single syntactic layer of representation and other frameworks that depend on more than one layer of representation. Here are some examples of theoretical frameworks of dependency that are represented by means of several layers of representations:

- Tesni`ere uses a single level of syntactic representation, the so-called ***stemma***, which on the other hand includes junction and transfer in addition to syntactic connection.
- In FGD Functional Generative Description there are two layers; an analytical layer, which can be characterized as a surface syntactic representation, and a tectogrammatical layer, which can be characterized as a deep syntactic (or shallow semantic) representation.
- In MTT Meaning-Text Theory recognizes both surface syntactic and deep syntactic representations (in addition to representations of deep phonetics, surface morphology, deep morphology and semantics).
- The framework of XDG Extensible Dependency Grammar can be seen as a compromise in that it allows multiple layers of dependency-based linguistic representations but requires that all layers, or dimensions as they are called in XDG, share the same set of nodes. This is in contrast to theories like

FGD, where e.g. function words are present in the analytical layer but not in the tectogrammatical layer.

A. Criteria for establishing dependency relations

Some of the criteria are syntactic and some are semantic. These criteria have been proposed for identifying a syntactic relation between a head H and a dependent D in a construction C [8], [9]:

- H determines the syntactic category of C and can often replace C.
- H determines the semantic category of C; D gives semantic specification.
- H is obligatory; D may be optional.
- H selects D and determines whether D is obligatory or optional.
- The form of D depends on H (agreement or government).
- The linear position of D is specified with reference to H.

8. EXISTING PARSING SYSTEMS

The most prominent dependency-based parsers are *Malt parser* and *Stanford Parser*. Malt parser is a data-driven dependency parsing system developed by Joakim Nivre. The system uses no grammar but relies completely on inductive learning from treebank data for the analysis of new sentences and on deterministic parsing for disambiguation [10]. The methodology of Malt parser is based on three essential techniques:

- Deterministic parsing algorithms for building dependency graphs.
- History-based feature models for predicting the next parser action.
- Discriminative machine learning to map histories to parser.

Stanford Parser is a statistical parsing system which provides both dependency analysis and phrase structure analysis for a set of languages including Arabic. The original version of this parser was written by Dan Klein. The current version of the parser requires Java 6 (JDK1.6).

9. DEPENDENCY TREEBANKS

A. The Quranic Arabic Dependency Treebank (QADT)

The Quranic Arabic Dependency Treebank is an annotated linguistic resource which shows the Arabic grammar, syntax and morphology for each word in the Holy Quran. The Quranic Treebank is mapping out the entire grammar of the Quran by linking Arabic words through dependencies. The linguistic structure of verses is represented using mathematical graph theory. The annotated corpus provides a novel visualization of Quranic syntax using dependency graphs. Fig. 5 shows an example of the dependency graphs of the QADT.

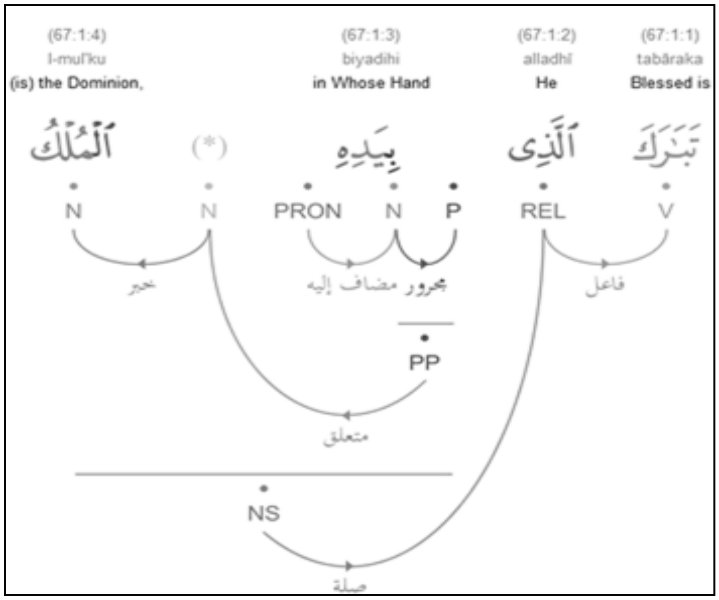


Figure 5: Dependency graph of the QADT

B. *The Penn Arabic Treebank (ATP)*

The ATB (2003) consists of 23,611 parse-annotated sentences [11] from Arabic newswire text in Modern Standard Arabic (MSA). The ATB annotation scheme involves 497 different POS-tags with morphological information (reduced to 24 basic POS-tags by Bikel e.g. NN, NNS, JJ), 22 phrasal tags e.g. NP, VP, PP and 20 functional tags e.g. SBJ, OBJ, TPC (52 combined functional tags, as functional tags can stack). The source text is a collection of newswire articles. Annotators use part-of-speech and phrase tags adapted from the English Penn Treebank project (over 400 tags are used). The grammar framework followed is constituent phrase structure grammar.

C. The Prague Arabic Dependency Treebank (PATB)

The PATB is the same collection of newswire articles, but annotated using a dependency grammar instead of using constituent phrase structure. The grammar framework used is a variation of dependency grammar called Functional Generative Description, originally developed at Prague in the 60's. The PADT 1.0 distribution comprises over 113500 tokens of data annotated analytically and provided with the disambiguated morphological information. In addition, the release includes complete annotations of Morphological Trees resulting in more than 148000 tokens, 49000 of which have received the analytical processing.

D. The Columbia Arabic Treebank (CATiB)

It is another re-annotation of the Penn Arabic Treebank newswire articles, but using a simplified dependency grammar which is closer to traditional Arabic grammar. The Columbia Arabic Treebank (CATiB) is a database of syntactic analyses of Arabic sentences. CATiB contrasts with previous approaches to Arabic tree banking in its emphasis on speed with some constraints on linguistic richness. A tagging scheme is used which allows rapid annotation. The Treebank uses only 6 part-of-speech tags, and 8 dependency relation types.

10. THEORETICAL FORMALISMS OF DEPENDENCY GRAMMAR

Dependency Grammar, like Phrase Structure Grammar, is just one approach in a theory of sentence structure. A wide range of more comprehensive theories include DG rather than PSG, and vary along much the same lines as theories which assume PSG: _ some reject transformations while others accept them. _ Some recognize a single level of syntax, while others disperse syntactic phenomena over a range of different levels of structure which map onto one another more or less freely. Some are sufficiently formalized to be used in computer systems, while others are relatively informal. _ Some insist on 'projectivity' (each word in a stemma 'projects' directly to its node, without crossing the projection line of any other word - i.e. phrases must be continuous), while others don't. The following list is a set of the most prominent theories of dependency grammar:

- Theory of structural syntax - Tesni`ere (1959)
- Word Grammar (WG)- Hudson (1984, 1990)
- Functional Generative Description (FGD) - Sgall et al., (1986)
- Dependency Unification Grammar (DUG) - Hellwig (1986, 2003)
- Meaning-Text Theory (MTT) - Mel'ˇcuk (1988)
- Lexicase - Starosta (1988)
- Constraint Dependency Grammar (CDG) - Maruyama, 1990, Harper and Helzerman, 1995; Menzel and Schr¨oder, 199)
- Constraint Dependency Grammar (WCDG) - Schr¨oder (2002)
- Functional Dependency Grammar (FDG) - Tapanainen and J¨arvinen, 1997; J¨arvinen and Tapanainen,1998
- Topological Dependency Grammar (TDG) - Duchier and Debusmann 2001
- Extensible Dependency Grammar (XDG) - Debusmann et al. (2004)
- Case Grammar - Anderson
- Daughter-Dependency Theory - Hudson
- Dependency Unification Grammar - Hellwig
- Functional-Generative Description - Sgall
- Metataxis- Schubert
- Unification Dependency Grammar - Maxwell
- Dependency Grammar Logic (DGL) - Kruijff, 2001

11. INVENTORY OF DEPENDENCY RELATIONS

Although most theories agree that dependency relations hold between lexical elements, rather than phrases, they can make different assumptions about the nature of these relations. In this research, a comparison was conducted between the inventories of the relations used in three systems in order to suggest the adopted inventory of dependency relations to be used in this research. The comparison was conducted between the relations of:

- Stanford Parser (53 relations)
- The Quranic Arabic Dependency Treebank (40 relations)
- The UNL system (38 relations)

12. DATA COLLECTION AND DATA ANALYSIS

There are two methods to collect data; Quantitative and Qualitative Data collection methods. Quantitative Data collection depends on random sampling and structured data collection instruments that fit diverse experiences into predetermined response categories. They produce results that are easy to summarize, compare, and generalize. Quantitative research is concerned with testing hypotheses derived from theory. Some of the quantitative data gathering strategies are:

- Experiments/clinical trials.
- Observing and recording well-defined events.
- Obtaining relevant data from management information systems.
- Administering surveys with closed-ended questions.

On the other hand, qualitative data collection plays an important role in impact evaluation by providing information useful to understand the processes behind observed results. Qualitative methods are characterized by the following attributes:

- They tend to be open-ended and have less structured protocols.
- They use triangulation to increase the credibility of their findings (i.e., researchers rely on multiple data collection methods to check the authenticity of their results)
- Their findings are not generalizable to any specific population.

Data collection in a qualitative study takes a great deal of time more than quantitative study.

The corpus that would be used in this research is the undiacritized corpus of the Arabic Language Technology Centre (ALTEC) [12]. This corpus was built to be one of the language resources for Arabic in order to support research in Natural Language Processing. The documents of ALTEC corpus itself were collected from the Arabic Wikipedia.

13. THE INTERACTIVE ANALYSER OF UNL

The Universal Networking Language (UNL) is an artificial language for representing information in a natural-language-independent format [13].In the UNL framework, there are two basic processes: UNLization and NLization. UNLization is the process of analyzing the information conveyed by natural language utterances into UNL. NLization is the process of generating a natural language document out of a UNL graph. The Interactive ANalyser (IAN) [14] is a natural language analysis system for the UNLization process; in other words, it is responsible for converting the natural language sentences into semantic networks in the UNL format.

In the UNL framework, the syntactic module is built on X-bar theory. X-bar theory may be not convenient for some languages such as Arabic, Irish and Welsh due to the free word-order of its structures. In Arabic we have the VSO structure that would violates the main schema of X-bar (SVO). In the VSO order there would be a specifier (the subject) between the verb and its complement. This may be solved by means of the process of Movement, but it still does not reflect the nature of the structures of that language. Form this point it would be more adequate to use the IAN system with its analytical capacity in tokenization, segmentation, transformation and disambiguation, and to reconsider the way of writing the analysis rules in a more convenient way for Arabic according to Dependency Grammar.

In its current release, IAN is designed to be parameterized to the source languages with a set of files:

- The input natural language document
- The NL-UNL dictionary (according to the UNL Dictionary Specs[15])
- The NL-UNL transformation grammar (transformation rules to convert natural sentences into UNL graphs (according to the UNL Grammar Specs[16])

- The NL-UNL disambiguation grammar.

Our target in this research is to generate a UNL network out of Arabic sentences according to Dependency grammar. So firstly it is necessary to discuss the components of the UNL network. The UNL network is composed of three main components; the Universal Words (UWs), the UNL relations, and the attributes. The UWs are the words of UNL to be interlinked by Universal Relations and specified by Universal Attributes - in a UNL graph.
Now we will have some examples for handling Arabic sentences in IAN. Regarding the sentence " أزمة فرنسا", the first step is to look up for the strings to get the corresponding UWs form the dictionary as shown in Fig. 6

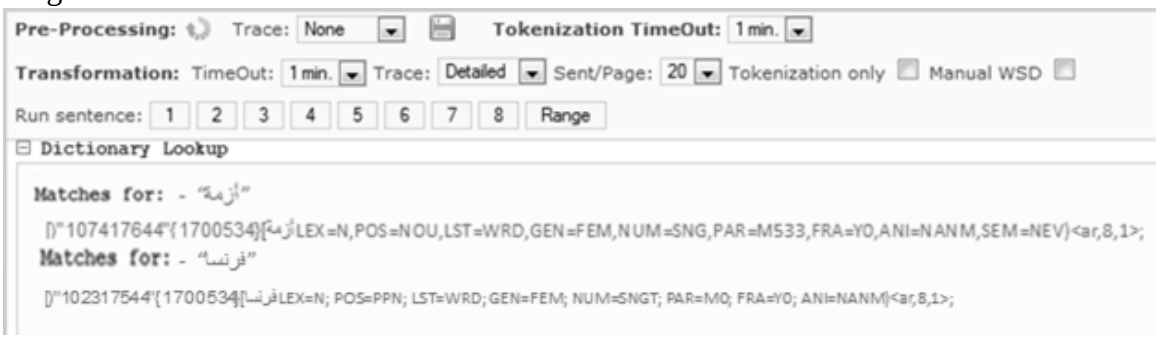


Figure 6: the dictionary look up for the string in the NL-UNL dictionary

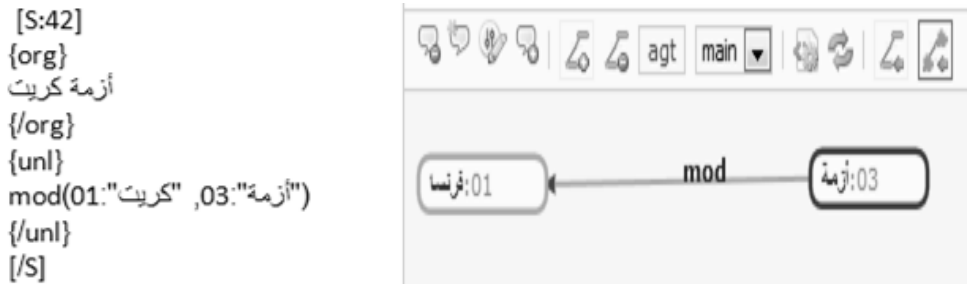


Figure 7: the UNL graph in IANFigure8: the nodes of the UNL graph

After adding the dependency rules that would tokenize and analyze that structure, the UNL graph would be as in Fig.7 and illustrated visually as in Fig.8. The rest of the other sentences of the selected corpus will be analyzed and processed.

REFERENCES

[1] Dick Grune and Ceriel J.H. Jacobs, *Parsing Techniques a practical guide*. Springer Science Business Media, LLC. 2008.
[2] Chomsky, Noam, *Syntactic Structures*. Gravenhage: Mouton. 1957.
[3] Chomsky, Noam, *Reflections on Language*. New York: Pantheon. 1975.
[4] Ryan McDonald, and Joakim Nivre, *Dependency Parsing: Synthesis Lectures on Human Language Technologies*.2009.
[5] Chomsky, N, *Aspects of the Theory of Syntax,* MIT Press, Cambridge, MA. 1970.
[6] Joakim Nivre. *Dependency Grammar and Dependency Parsing*. Växjö University. Sweden. 2005.
[7] Carroll, J, Statistical parsing, *Handbook of Natural Language Processing*, Marcel Dekker.2000.
[8] Zwicky, A. M. Heads. *Journal of Linguistics* 21: 1–29. 1985.

[9] Hudson, R. A., *English Word Grammar*. Blackwell. 1990.
[10] Joakim Nivre, *MaltParser: A language-independent system*. Natural Language Engineering 13. 2007.
[11] The Institute for Research in Cognitive Science websiteA. Bies and M. Maamouri: http://www.ircs.upenn.edu/arabic/Jan03release/guidelines-TB-1-28-03.pdf , (accessed September 2013
[12] Arabic Language Technology Center "ALTEC" website: http://www.altec-center.org/, (accessed September 2013).
[13] UNL website: http://www.unlweb.net/wiki/Introduction_to_UNL , (accessed September 2013).
[14] UNL website: http://dev.undlfoundation.org/analysis/index.jsp, (accessed September 2013).
[15] UNL website: http://www.unlweb.net/wiki/Dictionary, (accessed September 2013).
[16] UNL website: http://www.unlweb.net/wiki/Grammar, (accessed September 2013).

نحو بناء محلل نحوي للغة العربية المعاصرة

اعتماداًعلى نظرية التعليق

رحاب عرفات[١], سامح الأنصارى[٢]

قسم الصوتيات واللسانيات، كلية الآداب، جامعة الاسكندرية

الشاطبى، الاسكندرية، مصر

١rehab.arafat@gmail.com
٢sameh.alansary@bibalex.org

ملخص:

تهدف هذه الدراسة إلى عرض الخطة المتبعة عند بناء محلل نحوي للعربية المعاصرة اعتمادا على نظرية التعليق (نظرية التبعية). حيث يقدم نحو التعليق إطار لغوي مناسب للتحليل النحوي للجمل وخاصة للغات ذات النظام الحر في ترتيب الكلمات في بنية الجمل مثل اللغة العربية. وجدير بالذكر أن المبدأ الأساسي في نحو التعليق هو أن بينة الجملة تتكون من العناصر المعجمية يتم الربط بينها عن طريق علاقات ثنائية والتي تعرف بعلاقات التبعية. وتعود أصول نظرية التبعية إلى بعض النحاة العرب واللغويين اللاتينين وتم مناقشتها حديثا والرجوع إليها لعدد من الأسباب النظرية والتطبيقية. ويسعي هذا البحث إلى توضيح بعض القضايا الأساسية للتحليل النحوي استنادا على نظرية التعليق الوخطوات المتبعة في بناء محلل نحوي آلي بالاستعانة بالمحلل التفاعلي إيان (IAN) .

Building a Syntactic Parser for Arabic Verbal Phrases Based on X-Bar Theory

*Omnia Zayan*1, Sameh Al-Ansary**2*

**Phonetic and Linguistics Department, Faculty of Arts, University of Alexandria*

ElShatby, Alexandria, Egypt

[1]omnia.zayan@yahoo.com

***Phonetic and Linguistics Department, Faculty of Arts, University of Alexandria*

ElShatby, Alexandria, Egypt

***Bibliotheca Alexandria, Alexandria, Egypt*

[2]sameh.alansary@bibalex.org

Abstract— **This paper is concerned with presenting a linguistic description of structural and syntactic analysis of the verb phrases in Arabic based on X bar theory. The paper discuses the linguistic framework of X bar theory which claims that all languages share the same underlying syntactic structure. Then it goes to explain how to apply X bar on the Arabic verb phrases. The data gathered from the fairy tale "الأميرة النائمة". The tool that is used for modeling X-Bar Theory is a UNLization tool called the Interactive Analyzer (IAN) which follows the X bar approach.**

1 INTRODUCTION

The human mental representation of language is part of the linguist's task. This task can be achieved by constructing a model that represents the form and content of the human mental representation. The modular view of the mind means to construct a theory to explain the grammar of language. The relation between a theory of linguistic structure and particle grammar is that the theory must provide a practical and mechanical method for constructing the grammar given a corpus Chomsky (1957) [1]. Grammars should satisfy the following basic requirements: they should be observationally adequate, by being capable of demonstrating whether a particular string of words is well formed or not, also they should be descriptively adequate, by assigning structural descriptions to strings of well-formed sentences and they should be explanatorily adequate, by representing the best available descriptively adequate grammar of what kinds of grammars are possible for human languages [2].

X bar theory was first invented by Harris (1951) [3] but not by that name. The first presentation of X-bar theory appeared in Chomsky (1970) [4] and further developed by Ray Jackendoff (1977) [5]. There are two basic considerations that motivate the existence of X bar theory: phrase structure rules and cross-categorical structure [6]. According to Lyons (1968) [7] the first consideration fall to capture "Endocentricity" (the head node shares its categorical properties with the phrasal node containing it) which seems to be a fundamental property of human language. The second motivation for X bar theory to solve the problem of cross-categorical parallelisms by providing a generalized structure that express basic grammatical relations.

2 LITERATURE REVIEW

Many attempts have been proposed to explain Arabic syntactic analysis. Although X-bar structure is thought to be universal because it occurs in all languages, there are few attempts of syntactic analysis for Arabic based on X bar theory. Specific syntactic analysis using X bar model had been emphasized in this review: Al-Bayaty (1990) [8] provided a hypothesis that explains the Arabic negations structure within the framework of X-bar theory by discussing the two possibilities for c-selection: whether T c-selects NegP, or Neg c-selects TP. The proposed analysis of Al-Bayaty, depends on the assumption that Neg is a head in its own right and therefore interacts with V-movement. Kremers (2003) [9] reported results in the analysis of the Arabic noun phrase by giving an account for several of common phenomena known from Arabic noun phrases, such as the genitive construction, word formation,

placement of adjectives and other modifiers, adjectival agreement and the formation of deverbal nouns and participles. Another tries for applying X-bar theory was by Tamadla (2006) [10] who applied X-bar theory to the Arabic language in order to supply a systematic description of standard Arabic sentence formation. Tamadla reached a conclusion that X-bar theory is flexible enough to incorporate all cross-linguistic variation by putting different assumptions.

Finally, Most recently, Al Aqad (2013) [11] applied X bar theory on a multi-position Arabic adverbs to achieve the following objectives: offer a syntactic baseline between Arabic and English by comparing the Adverb position in both Arabic and English language, examines the adverbs function of Arabic language in the X' theory and to determine which syntactic theory is relevant to Arabic adverb. Six kernel sentences have been tested in his study, three sentences written in standard Arabic language and their correspondence in English, and three sentences written in Standard English and their correspondence in Arabic. Al Aqad concluded that the applications of X bar theory on Arabic language sentences show the differences and indiscernible constituents structures among languages and the basic that human are born with linguistic knowledge like what captured in X-bar theory.

3 LINGUISTIC FRAMEWORK

The name "X-bar theory" comes to indicate the intermediate categories. X-bar theory claims that all phrases in a sentence exist common similarities in all languages. All those languages share the same underlying syntactic structure.

A. Constituents of X bar Schema

X bar theory aims to identify the similar structures among languages. The three layers X, X' and XP are obligatory in X-bar schema as in figure (1).

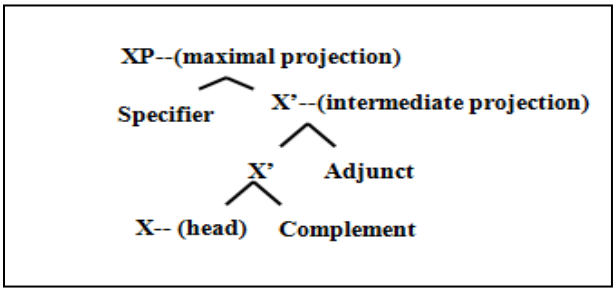


Figure 1: X bar schema

X (=X0) is the lexical projection of the vocabulary item that it dominates. Thus, the X may become an N for noun, a V for verb and so on. X' is the intermediate projection, and XP (= X'') is the phrasal or maximal projection of the head (which is also called a double-bar projection). XP includes X as the head, its complement, and its specifier. The head, the complement, the specifier and the adjunct are said to be the constituents of the syntactic representation and define the four general universal syntactic roles:

1) Head

The nucleus or the source of the whole syntactic structure which controls and determines the category of the other ones is called the head. All phrases must minimally contain a head and the constituents that follow the head are required to complete the sense of the head which means that the complement, adjunct and specifier branches in an X-bar structure are optional.

2) Complement, Specifier and Adjunct

The sister of the head is called its complement which is necessary to the head to complete its meaning. while the Specifier is an external argument appears to the left of X'. Adjuncts are always found as a

sister of bar-level categories in phrases and they are adjoined either to the right or to the left of single bar categories. There is an important difference between complement and Adjunct: complements are sisters of their head, while adjuncts are sisters of the single bar level above the head as illustrated above in figure 1. In verbal phrases structure adverbs are normally considered adjuncts because they do not be part of the argument structure of the verb but is added on to modify the meaning of the verbal phrase.

B. Advantages of X Bar Theory

1) Hierarchal Organization

X–bar theory provides principles for the projection of phrasal categories from lexical categories and imposes conditions on the hierarchical organization of categories in the form of general schema. According to this theory, the phrase structure component of human language consists of phrase structure rules that represent the hierarchal relation between the different projections which is reflected in terms of the number of bars associated with each projection, which are considered to be principles of Universal Grammar. This is shown in the following two basic phrase structure rules extracted from X bar schema in (1):

(1)

a. X'' → YP X'
b. X' → X^0 ZP

The hierarchy is from 'double bar' to 'single bar' to 'zero bar'. The double bar projection (X'') is referred to as the maximal projection, X' is called a single bar and X^0 is the head projection. YP in (1 a) stands for specifier which is a daughter of XP and a sister to X' and ZP in (1 b) stands for the complement of the head which is a sister of the head and a daughter of X'.

2) The Uniqueness of Mother Node

Another advantage of X bar grammar is the uniqueness of mother nodes which prevents a given element from being immediately dominated by more than one node as in (1 b) X^0 takes ZP for complement and both have the same mother node X' with no intermediate category.

3) Endocentricity

Endocentricity property means that the head node shares its categorical properties with the phrasal node containing it. It is a requirement in X-bar theory that phrases be endocentric. The endocentricity encoded in the X'-template thus emphasizes that all phrases have heads which determine their categorical nature (a noun project a noun phrase, a verb projects a verb phrase, etc.)

C. Binary Branching

Binary branching refers to the way in which sentences are derived and represented within tree diagrams which allows for every part of the tree diagram to diverge into two nodes with one head and one constituent. A problem for binary branching appears in the "double object structure" which contains a transitive verb that has two noun phrases (direct and indirect object) as its complement. It seems that taking the assumption that complements are sister to heads will solve this problem as in figure (2).

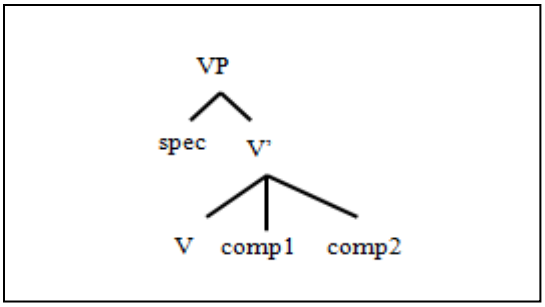


Figure 2: Complement as a sister to head

But in this traditional ternary-branching structure the V-bar branches out in to three separate constituents which is problematic because such analysis presumes that there are symmetries between the direct object and the indirect object. However, they are very different and need to be distinguished. Another problem is that there is no obvious way to deduce how either object receives the right case, or how the theta roles are assigned [12]. There are evidences that support the binary branching structure by reference to double-object structures: the first use of binary branching appeared in Ross's work Ross (1967) [13] which he called "Chomsky-adjunction" and then Kayne (1983) [14] proposed an "unambiguous path" condition on various syntactic relations and according to this condition branching are almost binary and then Larson (1988) [15] in which binary branching is considered the core of the VP-shell hypothesis.

On the other hands, there are opposition to binary-branching that reject the evidences that support binary-branching such as Simpler Syntax Hypothesis headed by Culicover and Jackendoff (2005) [16] and coordination structure because the arrangement of the categories is not hierarchical and there is no constituent can be established as the unique head of the conjunction. These exceptional nature of coordination are illustrated by multi-branching as in figure (3). This analysis has been proposed by many, Pullum and Zwicky 1986, Sag (1994) [17] and Ingria (1995) [18].

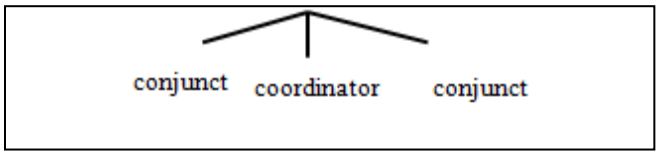


Figure 3: Multi-branching analysis of coordination

D. Extending the X-bar Schema to Functional Categories

Transformational grammar assumes that clauses are built up from sentences using the rule: S' → COMP S. This rule means that the head of the clause is the sentence and the complementizer is a specifier. According to X bar theory the sentence cannot be the head of any phrase but it is a complement. If the complementizer is the head of the clausal complement, then according to X-bar theory the clausal complement is a complementizer phrase (CP). Further, a recent version of Chomskyan theory (1986) [19] brought the non lexical (functional) categories into line with the lexical one so that the non lexical categories are also governed by the principles of X-bar theory. Much research on syntactic projection takes the view that projection is symmetric across syntactic categories. According to this view, the way in which functional information is mapped onto syntactic structure is fundamentally the same. This functional categories include: I/INFL, D/Det, Neg/NegP and C/Comp:

1) IP (Inflection phrase)

The category I/INFL stands for inflection. These inflections carry information about time, aspect, voice and tense. Tense is obligatory in all main clauses which always present in finite clauses even if it is not phonologically realized. IP is the maximal projection as in figure (4). In IP the head of the sentence is I, the complement of an IP is the predicate of the sentence VP and the specifier of an IP is the subject of the sentence.

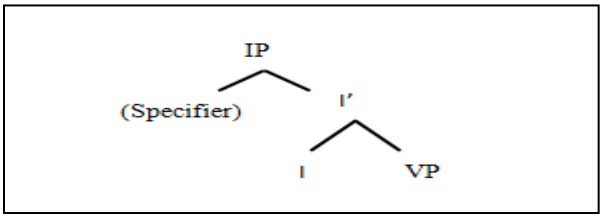


Figure 4: Inflection phrase (IP)

2) CP (complementizer phrase)

The idea of complementizer category was first proposed by Peter S. Rosenbaum in 1967 [20]. A CP is a phrase headed by a complementizer which is added to the beginning of an IP like subordinate clause. C is the syntactic head of embedded clauses. A complement of C is an IP and a VP is the full predicate which can be represented in figure (5).

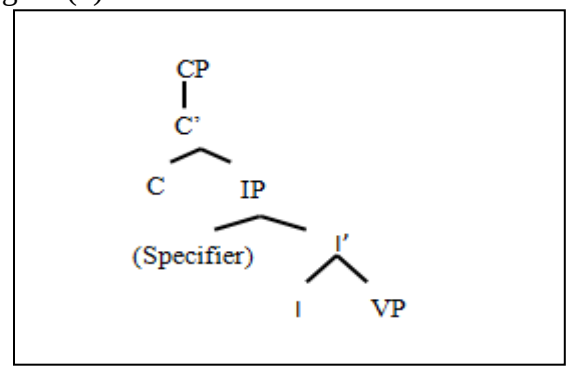


Figure 5: Complementizer phrase(CP)

This structure shows the complementizer in the C position of the CP is the head of the clause that follows. Some analyses allow for the possibility of empty complementizer which is represented by covert null category which is parallel in function to that of visible complementizer.

3) DP (Determiner phrase)

The category D/Det includes articles, demonstratives, possessives and quantifiers. Jackendoff (1977) assumes that determiners are specifier and the noun is the head. Jackendoff's analysis is an NP analysis in which the determiners are analyzed as specifiers. On the other hand Abney (1987) [21] proposed the DP analysis. DP analysis regards the structure of NP that has D in its specifier is problematic for X bar theory because specifiers are considered to be a position that host maximal projections and D in NP analysis is the only category which doesn't form a maximal projection and thus fall out of X bar schema. If D heads a DP, the DP will be the maximal projection of a determiner and the NP will be inside the DP as a complement to D. The difference between DP and NP analysis represented in figure (6).

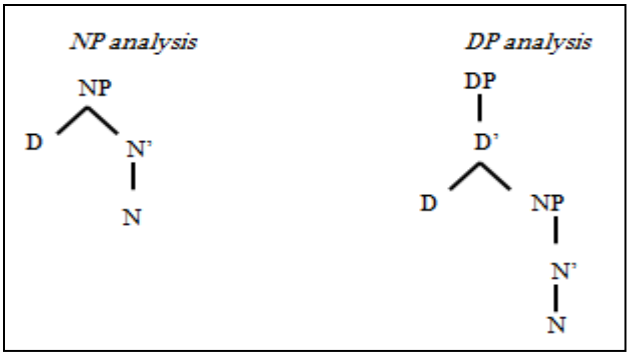


Figure 6: DP and NP analysis

DP analysis solves the problem of the non-phrasality of the D, But when more types of functional head are added, the nominal phrase no longer has a unique categorical definition.

4) NegP (Negation phrase)

In Chomsky's more recent work further functional phrases are posited. A negation phrase NegP, headed by a negation particle. Negation has its own projection which is placed in the tree between IP and VP as in figure (7).

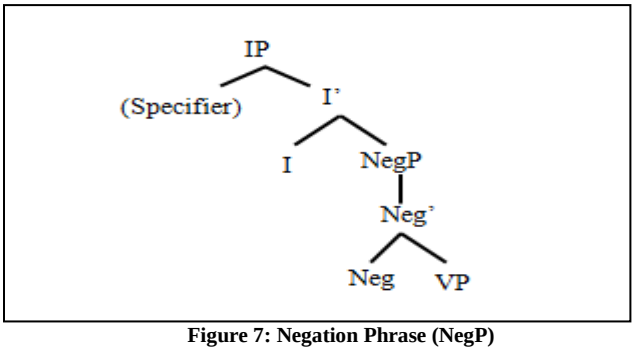


Figure 7: Negation Phrase (NegP)

E. Universality of X bar Theory

According to X bar theory no specifier or adjunct can intervene between the complement and the head but there are quite a number of languages for which the basic word order is Verb-Subject-Object (VSO) such as Irish, Welsh, and Arabic. In those languages the subject (a specifier) intervenes between the verb and the object so that X-bar theory cannot draw the tree for this structure. Chomsky (1957) observed that X bar theory cannot generate all the sentences of a language. He proposed that a set of rules that change the structure was needed. These rules are called transformational rules. Transformations take the output of X-bar rules (applied to the underlying structure) and change them into different trees. The output of a transformational rule is called the S-structure of a sentence. These transformation rules include both movement and insertion rule. Transformation rules will generate sentences that X bar theory itself cannot produce. For example, in the VSO word order X bar rules will be applied to its underlying order SVO which then by transformation rules (movement rules) turn to the surface structure (VSO).

4 METHODOLOGICAL FRAMEWORK

The researcher will assume only X bar theory restricted to binary branching in accordance with Chomskyan grammar in all structures except in the representation of coordinated constructions, because the binary branching cannot express the idea that the coordinated elements are at the same level. So, the researcher have decided to adopt the ternary branching for coordination (and only in this case) which gives the sentences depth, flexibility and grammatical accuracy. The study focuses on the analysis of the Arabic 'verb phrases' appear at the IP structures found at the surface structure. The researcher chooses the fairy tale "الأميرة النائمة" in order to analyze the verb phrases existing in it using the Interactive Analyzer (IAN) tool.

5 CORPUS DESCRIPTION

The data gathered from the fairy tale "الأميرة النائمة". This story can be said to be short enough to allow for UNLization which afford the possibility of generalizing the UNLization strategies to other similar texts. Additionally, the story offers the chance of experimenting the parsing of the Arabic verb phrases easily away from complexity.

Firstly, the text was manually segmented. The main processes in the text segmentation are determining sentences and word forms. Sentences generally end with known punctuations marks such as ".", "!", "،","?". The result of the segmentation is a corpus with the following characteristics: in a corpus of 1268 words making 124 sentences, there are only ٢7 sentences (at surface structure) have ٣0 verb phrases inside them, representing 2١.7 % of the total. The structures of those Arabic verb phrase are found in five types as in table 1: head alone, pre-head string and head, head and complement, head and adjunct and finally combinations of two or more of the previous.

TABLE 1
ARABIC VERB PHRASE STRUCTURE INSIDE IP

IP					
Spec	IB				
	I	VP			
		VB			Adjunct
		Pre-Head	Head	Complement	
الملكة			تبكي وتنتحب		
الأميرة		لن	تموت		
الخادمة			تنتف	الفروج	
الأمير			رأى	الذباب	نائما على الجدران
الأميرة	سوف		تنخز	إصبعها	بمغزل

6 INTERACTIVE ANALYZER (IAN) TOOL

The tool that the researcher will use for modeling X-Bar Theory is IAN. The UNDL Foundation built a UNLization tool called the Interactive Analyzer (IAN).The Interactive Analyzer (IAN) follows the X bar approach, it postulates that all human languages share the same underlying syntactic structure. It is a tool that is designed on linguistic background; taking into consideration the linguistic issues facing any

tool dealing with natural language texts. It includes a grammar for natural language analysis and operates semi-automatically. It is language-independent and the syntactic processing is done automatically through the dictionary and the natural language analysis grammar which is provided as separate interpretable files, but syntactic ambiguities are up to the user, who may backtrack and choose different syntactic paths. IAN tool exhibits enormous flexibility and opportunities in handling natural language text due to the fact that it is uniquely designed upon linguistic framework. Figure (8) shows the IAN tabs: NL input tab to provide the natural language document to be UNLized either by creating a new file or uploading an existing file, Dictionaries tab to provide the NL-UNL dictionaries, T-rules tab to provide the NL-UNL transformation grammar, D-rules tab to provide the NL-UNL disambiguation grammar, IAN console tab where the user will get the results. The IAN console brings the list of sentences appearing in the NL input, which may be processed one at a time, or in a range and the final tab is compare tab to compare the saved sentence with the result.

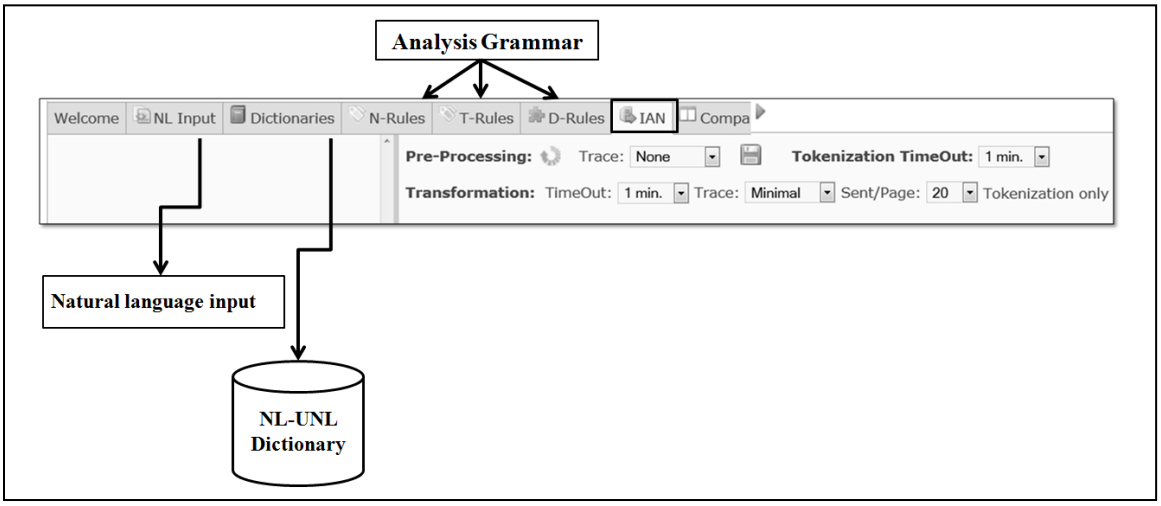


Figure 8: IAN tabs

7 A PRACTICAL EXAMPLE

In the UNL framework, a grammar along with dictionary constitute the basic resource foe UNLization. In order to use the tool, the user will have to sign in the UNL web at (http://www.unlweb.net/unlweb/) then access the IAN tool via UNL dev application. This section will present step by step the syntactic analysis of the Arabic verb phrase "تنخز إصبعها بمغزل" inside the IP "الأميرة سوف تنخز إصبعها بمغزل" using IAN tool. The UNLization process is performed in three different steps:

A. Tokenization

Tokenization is the process of segmenting the input into nodes. The system tries to match the strings of the natural language input against the entries existing in the dictionary. The tokenization process may be controlled by the NL-UNL D-Grammar as in figure (9).

Figure 9: Tokenization process

B. NL-UNL Grammar

1) List processing

These rules are used for pre-editing the natural language sentence and preparing the input to the syntactic module:

TABLE 2
LIST TO LIST RULE

Action	Rule	Description
Delete	(PUT,%x):=;	Delete punctuation
Replace	(POD,3PS,FEM,SNG,^@3,^@singular,^@female,^OPRON,%z):=([[00]] ,+@3,@singular,@female,+OPRON,+POSS,%z);	Replace the pronoun "ها" with the UW "00.@3.@singular.@female"

2) Intermediate projection

These rules are used to parse the head or any of its intermediate projections with complements (as in table 3) and adjuncts (as in table 4)

TABLE 3
INTERMEDIATE PROJECTION (XB): COMPLEMENTATION

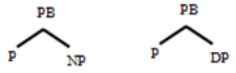
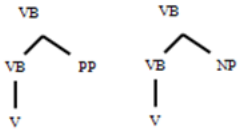
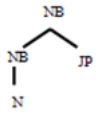
Action	Rule	Description	X bar representation
Complement of V	(V,%vb)({NP|DP|PP},%xp):=(VB(%vb,+proj,%xp,+comp,+proj),+XB=VB1,+LEX=V,%new);	Build the tee of "complement of the verb" by adding the head of the tree "V "to the NP or DP or PP tree	
Complement of I	(I,%ib)(VP,%right):=(IB(%ib,+proj;%right),+XP=IB,VP,+LEX=I,%new)(%rightt);	Build the "IB" tree by adding the auxilary to the maximal projection VP	
Complement of P	(P%ib)({NP|DP),%right):=(PB(%ib,+proj;%right),+XB=PB,+LEX=P%new)(%rightt);	Build the intermediate projection "PB" by adding the preposition to the maximal projection NP or DP	
Complement of N	(N,%nb)(DP,%dp):=(NB(%nb,+proj;%dp),+XB=NB,+LEX=N,%new);	Build the intermediate projection "NB" by adding the noun to the maximal projection DP	

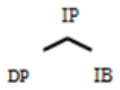
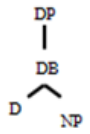
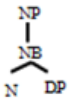
TABLE 4
INTERMEDIATE PROJECTION (XB): ADJUNCTION

Action	Rule	Description	X bar representation
Adjunct of the VB	(VB1,%vb)({PP|NP},%xp):=(VP(%vb,VB,%xp,+adjt,+proj),+XB=VP,+LEX=V,%new);	Build the intermediate projection VB by adding the first "VB" to the preposition phrase or to the noun phrase	
Adjunct of the NB	(NB%ib)(JP%right):=(NB(%nb,+proj;%jp,+adjt,),+XB=NB,+LEX=N,%new);	Build the intermediate projection NB by adding the NB to the maximal projection JP	

3) Maximal projection (XP)

These rules are used to combine the topmost intermediate projection and the specifier as in table 5:

TABLE 5
MAXIMAL PROJECTION (XP)

Action	Rule	Description	X bar representation
maximal projection IP	({DP|NP},%np)(IB,%vb):=(IP(%vb,+proj;%np,+spec,+proj),-IB,-VP,+XP=IP,+LEX=I,%new);	Build the tree of maximal projection IP by adding the specifier "DP" to the "IB" tree	
maximal projection DP	(D,%pb)(NP,%xp):=(DP(%pb,+proj;%xp,+comp,+proj),+XB=DP,+LEX=D,%new);	Build the tree of the maximal projection DP by adding the determiner to the "NP"	
maximal projection NP	(N%nb)(DP,%dp):=(NP(%nb,+proj;%dp,+adjt,+proj),+XB=NP,+LEX=N,%new);	Build the tree of "NP"	
maximal projection PP	(P,%pb)(NP,%xp):=(PP(%pb,+proj;%xp,+comp,+proj),+XB=PB,+LEX=P,%new);	Build the "preposition phrase" tree of by adding the preposition to the "NP"	

C. UNL Graph

After applying the previous rules, the UNL graph in figure (10) is formed.

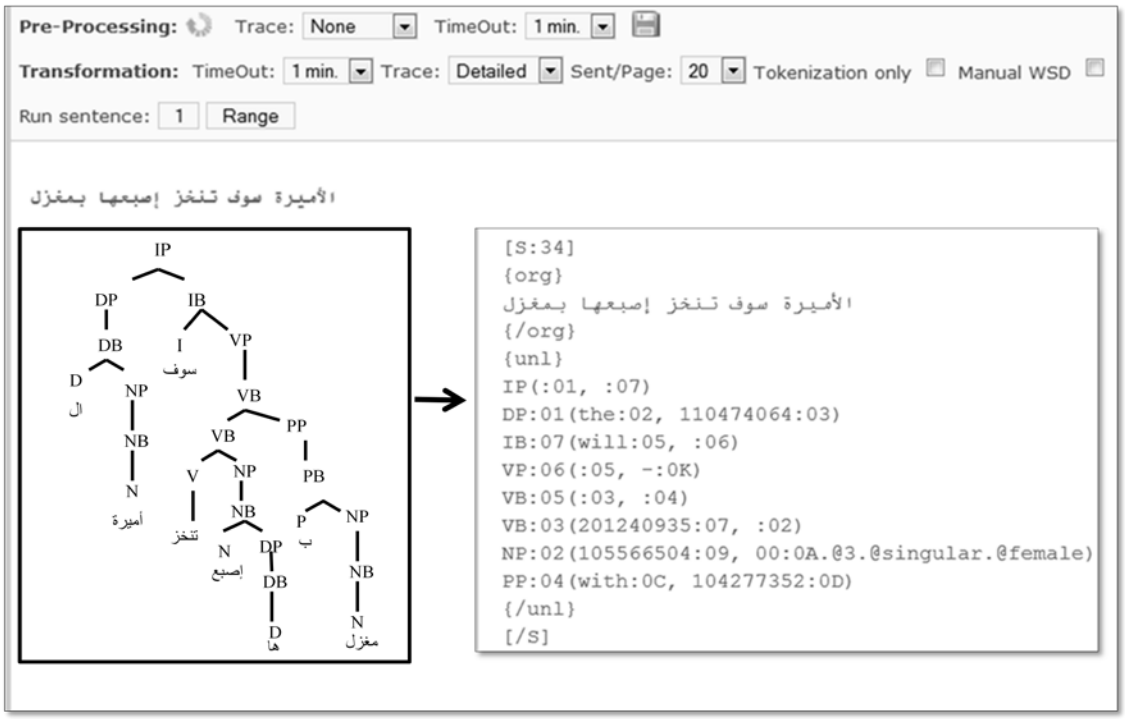


Figure 10: UNL graph

In figure 10, the VP built by adding (:03) to (:04). (:03) represents the first VB "تنخز إصبعها" (V + complement) and (:04) is the adjunct PP "بمغزل". When the auxiliary "سوف" added to the VP (:05), the IB is formed in (:06). The final stage is the addition of the specifier "الأميرة" (:01) to the IB to constitute the IP.

8 CONCLUSION

This paper presents a parser that explains the Arabic verb phrases structure within the framework of X-bar theory. The parser transforms input VPs into a syntactic tree in the UNL format. This VPs are gathered from the fairy tale "الأميرة النائمة". X bar provides a precise, flexible, computationally tractable representation for parsing.

References

[1] Chomsky, Noam. (1957). Syntactic Structures, The Hague/Paris: Mouton.
[2] W.J. Hutchins and H. L. Somers. (1992). An Introduction to Machine Translation Academic Press: London.
[3] Harris, Zellig S. (1951). Methods in Structural Linguistics, Chicago: University of Chicago Press.
[4]Chomsky, Noam. (1970). Remarks on nominalization. In R. Jacobs and P.Rosenbaum (eds.), Readings in Transformational Grammar. 184-221. Waltham, MA: Ginn.
[5] Jackendoff, Ray. (1977). X-bar-Syntax: A Study of Phrase Structure. Cambridge, MA: MIT Press.
[6] Baltin, Mark R., and Chris Collins, ed. (2001). The handbook of contemporary syntactic theory. Oxford: Blackwell.
[7] Lyons, John. (1968). Introduction to theoretical linguistics. Cambridge, England: Cambridge University.
[8] Al-Bayaty, Jumah. (1990). The X-Bar Status of NEG. Simon Fraser University.
[9] Kremers, J. (2003). The Arabic noun phrase: a minimalist approach. PhD Dissertation. Nijmegen: Catholic University.
[10] Tamadla, Said. (2006). X-bar Theory and Standard Arabic. Falun.
[11] Mohammed H. Al Aqad. Syntactic Analysis of Arabic Adverb's between Arabic and English: X Bar Theory. *International Journal of Language and Linguistics*. Vol. 1, No. 3, 2013, pp. 70-74.
[12] Stradmann, M. (2013). Reviewing the Binary Branching Hypothesis, BA thesis, Utrecht University.
[13] Ross, J. R. (1967). Constrains on Variables in Syntax, Doctoral Dissertation, MIT, Cambridge, MA.
[14] Kayne, R. S. (1983). Connectedness and Binary-branching. Dordrecht: Foris Publications.
[15] Larson, R. K. (1988). On the double object construction. Linguistic Inquiry, 19(3):335–391.
[16] Culicover, P. W. & Jackendoff, R. (2005). Simpler Syntax. New York: Oxford University Press.
[17] Pullum, Geoff, and Arnold Zwicky. (1986). Phonological Resolution of Syntactic Feature Conflict. Language 62:751–73.
[18] Ingria, Robert. (1995). The Limits of Unification. Proceedings of Association for Computational Linguistics, 194–204.
[19] Chomsky, N. (1986a). Barriers, Cambridge, Mass: The MIT Press.
[20] Rosenbaum, P. (1967). The grammar of English predicate complement constructions. Cambridge, Mass, M.I.T. Press.
[21] Abney, Steven Paul. (1987). The English Noun Phrase in its Sentential Aspect, Master's thesis, MIT.

بناء محلل نحوي للمركبات الفعلية العربية

(X-Bar)اعتماداً على نظرية

أمنية زيان[١], سامح الأنصارى[٢]

قسم الصوتيات واللسانيات، كلية الآداب، جامعة الاسكندرية

[١] omnia.zayan@yahoo.com
[٢] sameh.alansary@bibalex.org

ملخص_ هذا البحث يقدم تحليل لغوي على مستوى البنية النحوية للعبارات الفعلية داخل اللغة العربية مبنية على نظرية."X-Bar" يناقش هذا البحث الإطار اللغوي لنظرية "X-Bar" والتي تفترض وجود نفس البنية التحتية في جميع اللغات. أيضا طريقة تطبيق هذه النظرية على العبارات الفعلية في اللغة العربية سيتم شرحها ورسم الأشجار النحوية لها. العبارات الفعلية المستخدمة في التحليل تم تجميعها من قصة "الأميرة النائمة". الأداة المستخدمة في التحليل هي "IAN" وهي أداة لغوية تقوم بعملية تحويل اللغات الطبيعية إلى لغة "UNL" وتعمد في بنائها على نظرية "X-Bar" .

Semantic and Associative Relations in the Mental Lexicon: Evidences from Semantic Priming

Noha Fathy[*1], Sami Boudella[**2], Sameh Alansary[*3]

[*]*Phonetics and Linguistics Department, Faculty of Arts-Alexandria University*
noha_phonetics@yahoo.com
Sameh.alansary@bibalex.org

[**]*Faculty of Humanities and Social Sciences, United Arab of Emirates University*
s.boudella@uaeu.ca.ae

***Abstract*— This study examines a longstanding issue in the psycholinguistic literature; the semantic priming effect in which semantically related pairs induce shorter response time (RT) than unrelated ones. A set of semantically related and associated pairs were tested in three different prime durations; 100, 250 and 750ms using a standard lexical decision task, the results showed an automatic activation for the associated pairs which showed a facilitation effect at short durations (100, and 250ms) however the purely semantic pairs showed a facilitation effect at short and long prime durations (250 and 750ms) which specifies that such kind of relation involves both automatic and conscious processing for short and long duration respectively. The results are addressed in terms of the parallel distributed processing framework[1] and the spreading activation model[2].**

1 INTRODUCTION

Semantic priming paradigm is usually used to evaluate the implicit memory attention[3] as well as examining word recognition [4]. The semantic priming effect is an improvement in performance due to prior exposure to a related context, a target is facilitated due to presenting a related stimuli (prime) however the target may be inhibited in case of unrelated context. The issue of semantic priming has long been studied since [4] who made the concept of semantic priming comes into reality. The researchers claimed that speakers respond faster to a target when it is preceded by a semantically related word than when it is preceded by an unrelated word. Their work was first introduced in the Journal of Experimental Psychology entitled "Facilitation in Recognizing Pairs of Words: Evidence of Dependence between Retrieval operations." In their experiment, participants were tested using lexical decision task (LDT). They asked participants to decide whether two simultaneous strings were words. Half of the words were semantically related while the other half was not, they results showed faster response time for the related condition. Many subsequent studies following [4] observed the same pattern of results [5-9].

There are two attentional modes of priming as introduced by [3], who Manipulated four variables in a LDT; semantic relation, conscious expectation, shift and non-shift from one category to another and SOA, the experimenter introduced five kinds of pairs that manipulated the former four variables like when a bird is used as a prime a type of bird is expected (related, non-shift, and expected) however when a part of the body is presented as a prime a part of building is expected (unrelated, shift, and expected), [3] found that at short SOA related pairs showed facilitation whereas unrelated pairs showed inhibition, however at long SOA expected pairs showed facilitation whereas unexpected pairs showed inhibition. The researcher argued that there are two different processes operating at short and long SOA, at short SOA automatic facilitation is introduced whereas at long SOA a controlled processes operate. Controlled processing is also known as strategic processing, it is slow, sensitive to inhibition and facilitation [10]. Automatic processing is generally characterized by having a quick onset, occurs without intention and causes subsequent facilitation. Automatic processing of semantic priming is assumed to be the result of spreading of activation from one concept to another [2]. In priming experiments, automatic processing is manipulated through the use of short stimulus onset asynchrony (SOA) and low relatedness proportion (RP) whereas conscious processing is manipulated through the use of Long SOA and high RP. SOA is the time from the start of the prime to the onset of the target. Relatedness proportion is the percentage of related trials among all trials.

Semantic priming paradigm didn't only examine pairs that are related in terms of meaning like cat and dog but also pairs that are associated like bird and nest. Semantic priming is caused by words that are having common semantic features and bearing a semantic relation like antonymy, synonymy, category co-members, and hyponymy however associative priming is caused by the use of associated words. Associated words are produced in response to each other and they usually appear in the same context. Associated words are selected according to free association norms, like those of [11, 12]. In a free association task participants are asked to write the first word that comes into their mind in response to a given word. Associated words may or may not be semantically related. The previous discussion tabs three types of relations; associative semantic priming like cat and mouse, non-associative semantic priming which is usually called pure semantic priming like whale and dolphin, and non-semantic associative priming which is usually called pure associative priming like bird and nest. Associative priming is subject to directionality; association from the prime to the target like bell-hop is called forward association whereas association from the target to the prime like light-lamp is called backward association. Researchers didn't only examine association but they also examined association directionality [for review 13, 14-16]. It is empirically difficult to distinguish semantic and associative relations as stated by [17] "having

devoted a fair amount of time pursuing free-association norms, I challenge anyone to find two highly associated words that are not semantically related in some plausible way." At first the associative values of semantically related pairs were not controlled, in fact [4] semantic priming was actually associative priming; as the prime- target pairs were drawn from association norms rather than category norms. [18] was the first study to disentangle semantic and associative priming, he reported facilitation for both kinds of pairs however,[9] failed to replicate Fischler's finding and they argued that the semantic priming reported is due to conscious processing. Subsequent studies were equivocal in reporting semantic and associative priming [13, 14, 16, 19-23].

2 THE THEORETICAL FRAMEWORK OF SEMANTIC PRIMING

The automaticity of semantic priming is usually addressed in the framework of spreading activation model [2]. The spreading activation model assumes that concepts are represented by nodes which are joined together by links representing the relationship between concepts. In this model the organization is not hierarchical and the length of each link represents the degree of semantic relatedness between concepts. Information about words is stored in two separate networks; lexical network for storing phonological and orthographic information about words and a semantic network for storing concepts and those concepts are linked to the word forms in the lexical network. In the lexical network, nodes are connected to each other on the basis of phonological and orthographic similarity whereas in the semantic network nodes are connected to each other on the basis of semantic similarity. Connections between associated words exist at the lexical level while connections between semantically related words exist at the semantic level.

The spreading activation model assumes that when processing a concept, activation spreads to nearby concepts and that this activation level decreases as it moves outwards. For instance, red causes stronger activation of apple than sunset, because apple is closer in the network to red than sunset. When a second object is presented; the activation of the first concept is decreased. The model also predicts that only one concept is activated at one time due to the serial nature of human processing, but once activation occurs it spreads in parallel from the closer nodes to their associates.

Spreading activation model represents the traditional approach to semantic memory. On contrary, connectionist models [1, 24] represent the empirical approach to semantic memory investigation. Connectionist models are sometimes called parallel distributed processing or neural network models.

The model assumes that every node is connected to all the other nodes in the network in direct or indirect way in a recurrent neural network. Experiences and learning processes are a key function in adjusting the strength of connection among nodes; the model is constantly reshaped according to one's experiences. Connectionist models assume that concepts are represented by a pattern of activation across a network of interconnected units, the model assumes that similar concepts have similar pattern of activation. Connectionist models provide a reasonable explanation of semantic priming. When presenting a prime, the prime is processed until the network settles into a pattern (an attractor), when processing the target the network starts from the pattern of activation of the prime, the network will settle faster for related than for unrelated word, because the pattern of the related word is similar to the previous activated pattern.

[1] distributed model simulates semantic and associative relations. In this model, semantic relatedness among words is encoded by the degree of featural overlap in their semantic representation whereas associative relatedness is attributed to the frequency with which two words appear together during training. The model uses word understanding task, the abstract version used for simulation is meant to map written words to their meanings. The abstract representation of this model is shown in figure 1, the semantic representation uses eight different patterns over 100 semantic features, the eight patterns represent eight semantic categories and sixteen category exemplars are generated from each pattern, by changing some features. Eight of the exemplars were more typical than the other eight; the semantic representation was then randomly assigned orthographic representation, which consists of 20 orthographic units. The network consisted of 20 orthographic units which are totally connected to 100 hidden units which are totally connected to 100 semantic units; the semantic units are connected to each other and also to the hidden units.

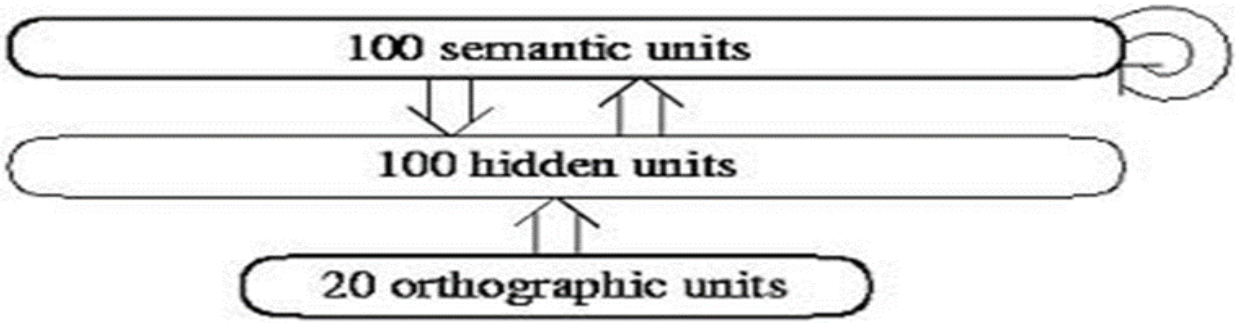


Figure 1: Plaut simulation of a simple task for mapping written words to their meaning [1]

The results of interest here are those related to associative and semantic priming. Regarding associative priming, Plaut found faster settling for associated than unrelated primes and the degree of priming was deeply affected by SOA and target frequency; greater priming effect was found at longer SOA and low frequency targets showed greater priming than high frequency targets. Regarding semantic priming the researcher found that semantic priming is much weaker and that

it is highly affected by target category dominance; more dominant category showed larger priming effect than less dominant ones. SOA had a different effect from that of associative priming; semantic priming effect is larger for short SOA and it gradually declines at longer SOA.

3 SEMANTIC AND ASSOCIATIVE PRIMING LITERATURE

Most of the studies over the semantic priming literature addressed both semantic and associative priming [19-22]. Whereas only few addressed both effects independently [5, 7, 9, 13, 14, 16]. In particular, the automaticity of both semantic and associative priming, the effect of various manipulations on the degree of priming effect including SOA, the presentation paradigm and the task. One of the earliest studies in examining pure semantic priming is [7], who used a set of semantically related pairs tested using LDT at various durations (360,600,2000ms), the results showed that the related condition showed a faster RT than the unrelated condition however, RTs tended to be faster at the 600ms prime duration. A more recent study is [5] who claimed that the degree of semantic relatedness influences the response latency, the stimuli used were a triplet of a dissimilar prime, similar prime and a target, 168 student participated in the experiment, the study showed that priming for highly similar items were found for short and long SOA, whereas priming for less similar items were only found for long SOA, which suggests that less similar pairs relies on some sort of strategic processing and priming for highly similar items is automatic. Generally speaking, semantically related pairs show shorter response time than their unrelated counterpart whereas highly related pairs affect the response time dramatically.

At the other end of the spectrum, very few studies addressed associative non-semantic priming [13-16]. One of the earliest studies in examining association directionality is [14], the study reported equal priming effect for forward and backward using the lexical decision task. The same results were replicated by [16, 25] and opposed to [26] who claimed that the reporting of backward priming is the result of semantic matching procedure. [16, 25] reported forward and backward priming in the LDT using 500ms prime duration while they failed to report backward priming in naming task. [14] examined compound and non-compound associates at two prime durations (150,500ms) in naming and lexical decision tasks, the results showed backward priming at the 150ms for both tasks, whereas at the 500ms it was reported only for the LDT. The previous discussion forces us to conclude that the reporting of backward associative priming is closely related to the task and the prime duration selected. LDT is capable of reporting priming effect at long and short SOA, whereas naming task reports backward priming only at short SOA, which suggests that LDT at long SOA relies on some sort of conscious processing.

A middle ground between the two is the examination of both effects simultaneously. [22] investigated pure associative and pure semantic priming, using a lexical decision task, three prime durations were tested (100, 250, and 500ms), the results supported the existence of pure automatic semantic priming as well as pure automatic associative priming but the average priming effect was larger for semantic priming than associative priming. Employing a different task, [19] examined semantic and associative priming in picture naming task and short SOA (114 and 234ms), the semantic pairs showed a significant inhibition at 114ms whereas the associated pairs showed significant facilitation at 234ms. The automaticity of semantic and associative relations was further examined by [9] using the single presentation procedure, the experimenters claimed that the single presentation procedure tabs automatic processing as participants are asked to respond to all the stimuli in sequence thus they stay unaware of the relation holding between pairs. The experimenters used a set of associated pairs that were also semantically related and a set of semantically related pairs. The results showed an associative facilitation in the single presentation condition which couldn't be found for the semantically related condition, thus stressing the automaticity of the associated relations and the presence of strategic processing for the semantically related pairs. In a more recent study, [23] reported semantic priming for pairs that are semantically related and unassociated for SOA ranging from 83 to 166ms using the LDT combined with the masked priming paradigm.

In the second experiment reported by [20], a set of pure semantically related as well as pure associated pairs were examined using a lexical decision task, the set of associated pairs were selected from [11] whereas the set of semantically related pairs were selected from the WordNet, the study showed a shorter response time for the associated pairs at 200ms prime duration.

In an analysis of the semantic priming literature, [27, 28] reviewed a set of studies that examined semantic and associative priming. [28] examination of twenty six studies revealed that there can be semantic priming without association but it is possible that the observed semantic priming is due to strategic rather than automatic processing, which is manifested through SOA greater than 250ms [10] and high relatedness proportion whereas no associative priming in the absence of semantic relation. Semantic priming effect is similar for different types of LDT whereas it is smaller for naming task. Adopting a similar perspective [27], investigated twenty four priming studies that are different from those of [28], [27] agreed with [28] upon the presence of associative boost and automatic priming for functionally related pairs like *hammer* and *nail* however [27] reported associative priming and criticized the priming reported by Lucas for the category coordinates as he claimed that it is due to including strategies that encourage strategic processing.

4 THE PRESENT STUDY

The experiment addressed in the present study focused on pure semantic priming without association and pure associative priming without semantic featural overlap. In order to achieve pairs of the former types, a pilot study of three

pretests was carried out. A set of semantically related pairs were selected from two studies that adopted similar approach [19, 29], these word pairs were then translated into Arabic using a standard English-Arabic dictionary [30] to be ready for the pretest phase. First a semantic categorization task that involved defining words on a semantic basis, through mentioning the major features marking it, for instance a dog is a four legged animal that barks. The aim of this task is to ensure the presence of a semantic relation between the prime and the target word, fifteen undergraduate students from the faculty of Arts Alexandria University were enrolled in this task, the words were presented randomly and intermixed with non-experimental words, the word pairs were judged as semantically related if they were mentioned in each other definition or defined with similar words and features. The second test was a similarity rating task following [31], in which participants are asked to rate the similarity between two words on a ten points scale. From 0 to 9, where 0 means no similarity and 9 means almost identical. The former test was applied to both kinds of pairs; semantic and associative, where thirty participants were involved.

Associated pairs are usually collected through free association norms [11, 12], in this study the target words previously reported in the semantic pairs were used to design a free association task. The targets were randomized and intermixed with non-experimental words, then presented to fifteen postgraduate students from the faculty of Arts, Phonetics and Linguistics department. They were then asked to name the first word that comes into their minds when reading each word, only the first response was encountered. The free association task resulted in an associated list that was further examined using a similarity rating task as reported previously in this section.

A. Participants

Forty five students from the faculty of Arts, Alexandria University volunteered in the study. Thirty in each duration.

B. Material

The word pairs in this experiment were a set of 39 semantic primes, 39 associated primes, and 39 baseline primes in addition to a set of 39 nonwords for using LDT see appendix1. The associated and the semantic pairs serve as the related condition and the baselines serve as the unrelated condition against which the degree of facilitation and inhibition is calculated, the non-words were constructed by changing one or two letters in the target words to form a pronounceable nonword. Further details were then specified for the primes and the targets; the pointed stem frequency, the root of each word were reported from ARALEX (A lexical database for Modern Standard Arabic) [32] to make sure no root is repeated twice, as well as specifying the orthographic ambiguity, the whole description of the stimuli is shown in table 1.

1) Calculation of the associative strength: The associative strength is calculated according to the number of participants from a population named a pair as associated [21, 27], The association strength was defined as one of the following; no association (have an average strength of less than 1%), weak association (have an average strength ranging from 1 to 10%), moderate association (have an average strength ranging from 10 to 20%) and strong association (have an average strength greater than 20%). Accordingly a pair is classified as strong associates if more than 20% of the participants would give the target as the primary response to the prime. In this study, the used population in the free association task was fifteen participants. After applying the previous criteria, associated pairs were classified into; 19.5% are weakly associated, 12.2% are moderately associated and 68.2% are strongly associated.

2) Calculation of the number of links between two concepts using WordNet: WordNet [33] provides us with several measures of similarity and relatedness[34]. Leacock and Chodorow (1998) similarity measure was selected to calculate the number of links between the related prime and target pairs. It is based on the path length between two concepts and it can be used to calculate the number of links holding between two concepts [31]. The pairs were first translated into English using a standard Arabic- English dictionary. WordNet 2.1 interface was first used to know the desired synset for each word, for instance for the word table, the second synset was selected (a piece of furniture) and so with the rest of the words. WordNet similarity was used as it provides a command line interface for each similarity type that can be written in a python shell to calculate the similarity between two concepts. The command line contained the following steps; the first step is meant to save the value of the first synset for the word *beauty* which is a noun, the second step is meant to do the same for the second synset. The third step is meant to calculate the Leacock and Chodorow (1998) similarity between the two synsets.

```
Beauty = wn.synset ('beauty.n.01')
Elegance = wn.synset ('elegance.n.01')
beauty.lch_similarity (elegance)
2.2512917986064953
```

C. Design

In this study, the stimuli were visually presented as the sensory memory for the visual stimuli is shorter than the sensory memory for the auditory stimuli [35]. The prime and the target were presented in a paired presentation procedure. Three types of primes were used, related primes which are either semantically related or associatively related and unrelated primes. Three lists were constructed for each SOA in a Latin square design (3×3) to form nine balanced lists where each contained thirteen semantic pairs, thirteen associated pairs, thirteen unrelated pairs and thirty nine nonword pairs, so that no prime target pair was repeated twice in the same list. The independent variables were the degree of

relatedness and the SOA (semantically related vs. associated vs. unrelated and SOA: very short (100ms), short (250ms), and long (750ms). The decision latency was set as the dependent variable.
As for the training set, participants were given eighteen training pairs, three of which were semantic another three were associative and three unrelated whereas the other nine were nonwords.

D. Procedures

Participants were tested individually; participants were seated approximately 60cm away from the monitor. Instructions and stimuli were presented on a 14 inch VAIO Intel core i5 using Superlab, participants were asked to read a set of instructions displayed on the monitor see appendix 2, which were then paraphrased orally by the instructor followed by the set of training pairs. After finishing the training phase, the experimenters made sure participants understood the procedures quite well, only at this stage participants were enrolled in the experimental phase. In the experimental phase, each participant was given a total of 78 trials, in which twenty six trials were related, thirteen unrelated trials, and thirty nine nonword trials.
Decision latencies were measured in milliseconds accuracy, using two keyboard buttons; / for word response and z for nonword response, participants used the dominant hand for the word response. Each trial consisted of a fixation cue (+), which was presented for 1000ms, followed by the prime which had a varying duration according to the SOA- the prime duration was 100ms, 250ms, 750ms in case when the SOA was 100ms, 250ms, 750ms respectively- the prime was then followed by an ISI of 50ms followed by the target which remained on the screen for 2000ms. For each trial participants were instructed to read the first word presented on the computer screen silently and to react to the second presented stimuli using yes and no response.

TABLE I

STIMULI CHARACTERISTICS

Condition	**Target**	**Semantic condition**	**Associated condition**	**Baseline condition**
Mean number of letters	3.7	3.9	3.6	3.9
Mean pointed stem frequency	76.5	95	132	101
Orthographic ambiguity	14 out of 39	8 out of 39	15 out of 39	10 out of 39
Number of links (WordNet)		2	1.14	
Similarity rating		4.7	6.36	
Associative strength			W: 19.5% M:12.2% S:68.2%	

5 RESULTS AND DISCUSSION

The mean decision latency was calculated for each prime type (semantic, associative and baseline) over the three prime durations (100, 250, 750ms). Mean lexical decision latencies and net priming are shown in table II. Generally, the study revealed that associative relations showed larger priming effect than the semantically related pairs which is consistent with [20], however both semantic and associative relations peaked at 250ms [22] as shown in figure 2. Semantic relations showed an inhibition effect at very short duration as shown in figure 3 as reported by [19] using picture naming task, however this can be attributed to the nature of semantic relations that involve the activation of many nodes and features to induce facilitation which couldn't be done at this short duration this view of semantic activation is consistent with the global assumption of the spreading activation model however it is opposed to [21, 23] who reported semantic facilitation at this short duration. On contrary, associative relations showed an inhibition effect at long duration which ensures that associative relations relies on automatic processing as reported by [9, 22, 27] however it is inconsistent with [22].
According to the local processing assumption of the spreading activation model, activation and processing of a node takes time and this activation is attenuated through time, however the duration of conceptual processing increases when this node is connected to a large number of nodes and vice versa and that explains why semantic relations showed inhibition at very short duration then peaking at 250ms and at last damping at

750ms but it still can maintain facilitation at this long duration, however associative relations in its theoretical definition gather words together based on usage however they do not involve a lot of common features like (*bird* and *nest*) thus having shorter processing time, accordingly showed a great deal of facilitation at very short duration. As opposed to this view, the distributed network model of Plaut [1] showed larger associative priming at long prime duration as well as larger semantic priming at short prime duration which is opposed to the previous reported results.

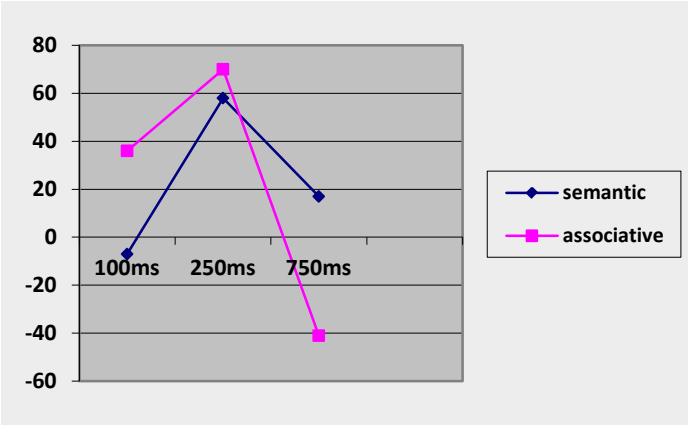


Figure 2: a curve specifying the effect of prime duration on both semantic and associative priming

TABLE II

MEAN DECISION LATENCIES

Prime duration	Semantic	Baseline	Net Priming
100ms	791	784	-7
	Associative		
	748	784	+36
250ms	Semantic	Baseline	
	693	751	58
	Associative		
	681		70
750ms	Semantic	Baseline	17
	751	768	
	Associative		-41
	809		

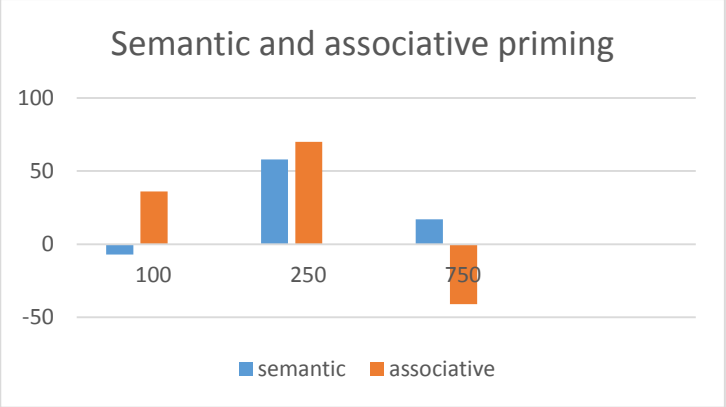


Figure 3 semantic and associative net priming

6 CONCLUSION

Semantic and associative relations proved to be quite different in terms of activation and processing. The presence of semantic facilitation at short and long durations forces us to conclude that semantic priming can rely on both automatic and conscious processing whereas the absence of associative facilitation at long duration ensures the automaticity of associative relation. Semantic relations involve more nodal activation than associative relations [for review 2, 36] which can be viewed in the absence of semantic facilitation at very short duration as opposed to associative relations, thus at the moment of speaking, examining the behavior of the reaction time across different prime durations seems the only way to figure more in-depth knowledge about the mental representation of various semantic relations.

REFERENCES

[1] Plaut, D.C. *Semantic and associative priming in a distributed attractor network.* in *Proceedings of the 17th annual conference of the cognitive science society*. 1995.
[2] Collins, A.M. and E.F. Loftus, *A spreading-activation theory of semantic processing.* Psychological review, 1975. **82**(6): p. 407.
[3] Neely, J.H., *Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention.* Journal of Experimental Psychology: General, 1977. **106**(3): p. 226.
[4] Meyer, D.E. and R.W. Schvaneveldt, *Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations.* Journal of experimental psychology, 1971. **90**(2): p. 227.
[5] McRae, K. and S. Boisvert, *Automatic semantic similarity priming.* Journal of Experimental Psychology: Learning, Memory, and Cognition, 1998. **24**(3): p. 558.
[6] Moss, H.E., et al., *Accessing different types of lexical semantic information: Evidence from priming.* Journal of Experimental Psychology: Learning, Memory, and Cognition, 1995. **21**(4): p. 863.
[7] Neely, J.H., *Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes.* Memory & Cognition, 1976. **4**(5): p. 648-654.
[8] Perea, M. and A. Gotor, *Associative and semantic priming effects occur at very short stimulus-onset asynchronies in lexical decision and naming.* Cognition, 1997. **62**(2): p. 223-240.
[9] Shelton, J.R. and R.C. Martin, *How semantic is automatic semantic priming?* Journal of Experimental Psychology: Learning, Memory, and Cognition, 1992. **18**(6): p. 1191.
[10] Posner, M. and C. Snyder, *Facilitation and inhibition in the processing of signals.* Attention and performance V, 1975: p. 669-682.
[11] Nelson, D.L., C.L. McEvoy, and T.A. Schreiber, *The University of South Florida free association, rhyme, and word fragment norms.* Behavior Research Methods, Instruments, & Computers, 2004. **36**(3): p. 402-407.
[12] Postman, L.J. and G. Keppel, *Norms of word association*. 1970: Academic Press New York.
[13] Kahan, T.A., J.H. Neely, and W.J. Forsythe, *Dissociated backward priming effects in lexical decision and pronunciation tasks.* Psychonomic Bulletin & Review, 1999. **6**(1): p. 105-110.
[14] Koriat, A., *Semantic facilitation in lexical decision as a function of prime-target association.* Memory & Cognition, 1981. **9**(6): p. 587-598.
[15] Peterson, R.R. and G.B. Simpson, *Effect of backward priming on word recognition in single-word and sentence contexts.* Journal of Experimental Psychology: Learning, Memory, and Cognition, 1989. **15**(6): p. 1020.
[16] Seidenberg, M.S., et al., *Pre-and postlexical loci of contextual effects on word recognition.* Memory & Cognition, 1984. **12**(4): p. 315-328.
[17] McNamara, T.P., *Semantic priming: Perspectives from memory and word recognition*. 2004: Psychology Press.
[18] Fischler, I., *Semantic facilitation without association in a lexical decision task.* Memory & Cognition, 1977. **5**(3): p. 335-339.
[19] Alario, F.-X., J. Segui, and L. Ferrand, *Semantic and associative priming in picture naming.* The Quarterly Journal of Experimental Psychology: Section A, 2000. **53**(3): p. 741-764.
[20] Buchanan, E.M., *Access into Memory: Differences in Judgments and Priming for Semantic and Associative Memory.* Journal of Scientific Psychology, 2009: p. 1-8.
[21] Bueno, S. and C. Frenck-Mestre, *The activation of semantic memory: Effects of prime exposure, prime-target relationship, and task demands.* Memory & Cognition, 2008. **36**(4): p. 882-898.
[22] Ferrand, L. and B. New, *Semantic and associative priming in the mental lexicon.* Mental lexicon: Some words to talk about words, 2003: p. 25-43.
[23] Perea, M. and E. Rosa, *The effects of associative and semantic priming in the lexical decision task.* Psychological research, 2002. **66**(3): p. 180-194.

[24] Plaut, D.C. and J.R. Booth, *Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing.* Psychological review, 2000. **107**(4): p. 786-823.
[25] Thompson-Schill, S.L., K.J. Kurtz, and J.D. Gabrieli, *Effects of semantic and associative relatedness on automatic priming.* Journal of Memory and Language, 1998. **38**(4): p. 440-458.
[26] Hutchison, K.A., *The effect of asymmetrical association on positive and negative semantic priming.* Memory & Cognition, 2002. **30**(8): p. 1263-1276.
[27] Hutchison, K.A., *Is semantic priming due to association strength or feature overlap? A microanalytic review.* Psychonomic bulletin & review, 2003. **10**(4): p. 785-813.
[28] Lucas, M., *Semantic priming without association: A meta-analytic review.* Psychonomic bulletin & review, 2000. **7**(4): p. 618-630.
[29] Sánchez-Casas, R., et al., *The nature of semantic priming: Effects of the degree of semantic similarity between primes and targets in Spanish.* European Journal of Cognitive Psychology, 2006. **18**(2): p. 161-184.
[30] Baalbaki, M., *A Modern English- Arabic Dictionary*. 2005, Beirut- Lebanon: Dar El-Ilm Lil-Malayen.
[31] Maki, W.S., L.N. McKinley, and A.G. Thompson, *Semantic distance norms computed from an electronic dictionary (WordNet).* Behavior Research Methods, 2004. **36**(3): p. 421-431.
[32] Boudelaa, S. and W.D. Marslen-Wilson, *Aralex: a lexical database for Modern Standard Arabic.* Behavior Research Methods, 2010. **42**(2): p. 481-487.
[33] Miller, G.A., *WordNet: a lexical database for English.* Communications of the ACM, 1995. **38**(11): p. 39-41.
[34] Pedersen, T., S. Patwardhan, and J. Michelizzi. *WordNet:: Similarity: measuring the relatedness of concepts.* in *Demonstration Papers at HLT-NAACL 2004*. 2004: Association for Computational Linguistics.
[35] Darwin, C.J., M.T. Turvey, and R.G. Crowder, *An auditory analogue of the Sperling partial report procedure: Evidence for brief auditory storage.* Cognitive psychology, 1972. **3**(2): p. 255-267.
[36] Anderson, J.R. and L.M. Reder, *The fan effect: New results and new theories.* JOURNAL OF EXPERIMENTAL PSYCHOLOGY GENERAL, 1999. **128**: p. 186-197.

Appendix 1

Target	**Associative-Prime**	**Semantic-Prime**	**baseline**
بغل	**عربة**	**حمار**	**نهار**
سرير	**نوم**	**أريكة**	**مفتاح**
خيط	**منبر**	**صوف**	**نار**
كأس	**عصير**	**فنجان**	**كوبرى**
سكين	**لحم**	**سيف**	**لعبة**
خروف	**عيد**	**ماعز**	**جراج**
عاصفة	**غبار**	**اعصار**	**محاسب**
طائر	**عش**	**باب**	**نسر**
حمامة	**سلام**	**نورس**	**حقيبة**
شباك	**نور**	**نسر**	**حكمة**
شارع	**مرور**	**طريق**	**خوذة**
سِنّ	**طفل**	**ناب**	**زجاج**
سوار	**حلية**	**خاتم**	**خيار**
خوخ	**بذرة**	**كمثرى**	**بنك**

سقف	**بيت**	**سطح**	**حلم**
سجادة	**صلاة**	**حصير**	**دراجة**
بطة	**بحيرة**	**وزة**	**دلو**
قرش	**دم**	**حوت**	**درس**
مَطَر	**شتاء**	**ثلج**	**زهرة**
قرد	**موز**	**غوريلا**	**سجق**
تفاحة	**حمراء**	**برتقالة**	**سحابة**
عنب	**خمر**	**زبيب**	**شهب**
زجاجة	**لبن**	**قارورة**	**شقيقة**
جزر	**أرنب**	**فجل**	**دقة**
ركبة	**رجل**	**مرفق**	**مربع**
دجاجة	**بيضة**	**كتكوت**	**شوربة**
ثعلب	**مكر**	**ذنب**	**خلاف**
قدم	**حذاء**	**يد**	**كتاب**
شاطئ	**رمل**	**ساحل**	**سفير**
قطة	**فأر**	**كلب**	**مسبح**
سفينة	**غرق**	**مركب**	**مصباح**
جنين	**أم**	**رضيع**	**شجرة**
ثوب	**فستان**	**رداء**	**سلك**
عطش	**ماء**	**جوع**	**سيدة**
دقيقة	**ساعة**	**وقت**	**صديقة**
جرذ	**مصيدة**	**سنجاب**	**نفط**
ثور	**مصارعة**	**بقرة**	**فرح**
منضدة	**طعام**	**طاولة**	**صنوبر**
بنت	**حسناء**	**ولد**	**عمق**

Appendix2

سنعرض عليك أزواجا من الكلمات العربية على شاشة الكمبيوتر. الكلمة الأولى من كل زوج ستكون مكتوبة بأحرف صغيرة و ستكون كلمة معروفة لديك مثل كلمة صحيفة أو طائرة. المطلوب منك أن تقرأ هذه الكلمة وأن تركز انتباهك على الشاشة حتى تظهر لك الكلمة الثانية من الزوج و التي ستكون مكتوبة بأحرف **كبيرة**. عندما ترى الكلمة الثانية الرجاء أن تضغط على الزر **"ظ"** في حالة التعرف على الكلمة و على الزر **"ئ"** في حالة عدم معرفة الكلمة. اعلم سلفا أن ٥٠% من الكلمات المكتوبة بالأحرف الكبيرة كلمات مختلقة و لا توجد أصلا في اللغة مثل **طقاب** أو **جليثة** في حين أن ٥٠% منها متداول و معروف لديك مثل **طاولة** أو **مدينة**

اضغط على أي زر لتبدأ التجربة. الرجاء أن تكون إجابتك سريعة و دقيقة.

Appendix 3: Experiment 1; 100ms

sem	796.2667	587	849	709	690	769	719	1092	681	720	1038	852	1081	683	608	866
sem	717.8	700	805	442	708	787	517	636	467	692	1071	798	904	865	480	895
sem	732.6667	604	732	687	606	545	640	738	532	594	1047	757	1175	828	589	916
sem	729.2	573	653	714	472	571	692	735	615	597	864	1129	1163	1035	492	633
sem	883.2	905	681	1021	544	706	995	698	755	610	922	1289	1704	772	566	1080
sem	826.6	1194	711	768	595	1001	659	824	635	950	887	1272	1049	719	424	711
sem	768.4	817	608	637	460	845	595	597	644	835	1068	989	1371	750	519	791
sem	773.2	629	939	782	485	649	827	855	740	640	538	831	1117	1326	572	668
sem	826.7333	706	695	500	1005	716	788	603	545	710	994	1051	978	745	624	1741
sem	824.2667	897	1089	1007	608	1204	480	723	558	789	991	924	800	852	499	943
sem	858.3333	898	1057	728	690	684	658	732	531	655	1012	1397	1534	638	567	1094
sem	808.5333	693	705	577	612	802	834	873	801	741	1005	1277	962	754	1019	473
sem	745.0667	746	698	550	546	1044	715	748	588	1098	0	1039	1325	708	524	847
	791.559															
asso	682.1333	743	788	734	535	568	920	597	717	797	550	864	0	951	576	892
asso	773.5333	776	744	613	644	597	1347	706	576	595	1006	988	1078	683	475	775
asso	782	759	606	518	1097	624	584	888	545	1003	601	800	1602	899	525	679
asso	664.9333	742	599	653	517	621	676	885	653	534	773	898	1241	646	536	0
asso	750.8	1087	782	580	572	750	693	668	697	615	687	1118	1109	734	577	593
asso	762.4667	974	539	579	548	571	729	924	892	713	777	735	1225	1125	575	531
asso	781.3333	617	632	552	543	707	963	817	756	654	675	1121	1637	733	611	702
asso	680.6	738	581	605	548	751	716	723	552	613	874	761	843	666	516	722
asso	722.6	644	525	534	535	824	648	773	482	663	814	1003	778	895	823	898
asso	704.2667	578	563	423	471	593	669	1072	703	949	1003	630	952	646	597	715
asso	813.4667	1247	849	479	852	817	693	1087	581	665	817	1184	1021	695	536	679
asso	705.9333	614	881	382	620	724	766	930	854	648	640	0	1021	0	866	1643
asso	901.2	614	546	489	468	745	681	840	591	648	720	1832	1864	1320	511	1649
	748.0974															
base	888.8	924	839	711	714	825	453	713	740	690	1866	925	1561	931	591	849
base	757.8667	759	750	496	521	677	556	657	600	963	1223	1290	999	787	500	590
base	736.0667	777	920	725	598	657	614	613	950	837	774	832	709	784	510	741
base	783.4	783	758	653	844	634	748	727	521	770	904	897	1561	749	498	704
base	727.4667	628	648	729	550	575	876	849	676	674	1158	630	699	951	516	753
base	691.2667	609	676	543	495	592	756	763	715	675	736	1140	0	905	795	969
base	797.6667	764	707	616	596	625	723	1083	585	657	877	692	1745	663	591	1041
base	746.6	549	612	761	664	914	443	751	570	686	877	835	1068	965	637	867
base	697.4667	585	542	656	618	577	498	967	585	824	831	786	1087	687	613	606
base	736.6667	647	704	566	539	554	529	1033	639	611	994	1181	1661	850	542	0
base	732.2	854	743	629	496	544	645	1009	495	685	912	673	789	673	534	1302
base	968.5333	946	903	1355	913	1258	489	670	623	761	996	740	1427	925	685	1837
base	938.6667	1110	844	1356	1003	1775	532	1301	601	670	1115	1081	845	708	444	695

Appendix4: experiment 2; 250ms

asso	711.8667	662	617	429	947	699	535	605	866	527	675	1238
asso	718.2667	1018	659	876	881	633	598	626	788	703	788	882
asso	625.8	649	542	499	586	565	681	534	748	515	666	633
asso	619.9333	562	640	663	634	575	516	564	928	490	649	668
asso	647	805	611	732	516	623	640	675	617	546	660	679
asso	751.4	633	603	539	866	435	572	600	1810	1048	935	970
asso	707.8	677	680	496	654	496	933	558	655	493	1445	981
asso	642.5333	536	803	414	643	449	679	647	1015	439	866	643
asso	647.7333	551	632	594	662	568	945	586	560	523	789	674
asso	726.4	816	419	510	838	475	960	885	820	611	1345	676
asso	677	755	1033	501	544	569	727	689	906	692	746	734
asso	746.6667	951	590	489	540	434	518	738	980	647	791	1011
asso	639.3333	487	499	510	557	449	540	533	741	597	888	1010
mean	681.6718											
base	808.7333	1359	631	741	787	902	1026	801	1089	585	972	763
base	665.9333	654	586	531	466	486	726	657	882	675	925	815
base	645.9333	668	611	528	568	720	615	532	1072	488	850	587
base	639.6667	623	568	526	588	530	603	558	786	586	698	932
base	700.6	1082	639	614	577	809	581	669	643	654	899	871
base	794	855	783	466	644	445	649	581	782	1177	814	1082
base	795.8667	1155	480	641	600	581	805	600	1080	549	1237	748
base	724.2667	825	446	531	827	521	569	602	986	641	847	559
base	725.8	618	420	1612	719	630	498	652	776	549	737	1091
base	750.0667	964	483	540	1767	596	534	644	776	495	563	1117
base	845.9333	584	557	558	743	605	902	618	1724	731	1555	792
base	954.1333	1554	731	757	1302	735	901	558	987	661	1576	855
base	721.9333	972	798	677	598	1261	551	544	612	1087	604	607
mean	751.759											
sem	719.8	1080	459	478	701	665	981	566	1049	630	953	913
sem	659.8667	595	530	513	696	404	635	560	1108	477	800	1162
sem	657.4	718	725	604	544	447	618	627	808	419	1262	666
sem	608.8667	778	513	539	646	375	693	627	551	505	744	923
sem	717.5333	1009	596	478	817	699	695	610	829	542	1030	622
sem	655.5333	1016	659	476	639	457	754	646	564	459	930	651
sem	668.1333	659	509	797	673	490	566	546	804	491	852	1308
sem	652.2	978	534	502	964	726	536	656	615	547	873	682
sem	651.0667	705	718	571	554	673	602	649	875	509	619	716
sem	843.8667	1420	768	1036	776	1306	662	513	1338	535	1125	687
sem	675.8	739	601	540	786	862	818	505	987	531	576	764
sem	847.1333	748	1043	700	1055	512	994	687	935	688	1347	1148
sem	659	517	518	589	535	564	565	666	0	575	1818	821
mean	693.5538											

Appendix 5: experiment 3; 750ms

asso	924	1603	1143	695	674	574	813	1553	502	1371	1175	494	850	655	944	814
asso	766.8	776	952	928	913	726	617	1180	537	724	780	550	658	494	909	758
asso	835.2667	1228	641	871	1264	564	1076	687	499	654	828	920	859	662	986	790
asso	698.2	804	968	773	866	529	555	609	577	558	735	648	792	652	815	592
asso	785.1333	1400	809	759	631	618	808	818	533	745	1020	670	687	622	811	846
asso	657.0667	832	573	731	641	596	697	972	552	686	0	598	657	781	964	576
asso	799.4	970	956	686	581	636	538	1357	787	979	671	645	967	646	900	672
asso	801.7333	858	1251	950	475	629	658	816	724	866	957	669	891	799	597	886
asso	910.3333	997	723	885	1380	589	769	1803	756	499	1566	629	670	483	1124	782
asso	851.7333	704	493	1190	1352	586	1003	903	936	898	979	607	792	625	810	898
asso	770.7333	872	570	670	1916	648	781	1054	485	657	662	633	812	484	894	423
asso	944.9333	1007	839	750	862	603	858	1006	690	1082	892	776	1238	908	1637	1026
asso	779.3333	705	841	472	553	549	861	610	466	1400	670	725	839	621	1373	1005
mean	809.5897															
base	738.1333	1072	942	943	822	494	756	688	607	781	708	532	726	467	752	782
base	822.8	998	755	674	1360	765	992	700	614	865	992	612	769	580	752	914
base	669.0667	765	615	669	580	863	657	747	460	723	715	632	690	638	701	581
base	667.7333	709	585	663	542	565	475	644	555	1109	679	489	611	726	868	796
base	736.6667	753	1058	656	789	602	771	838	628	691	670	626	730	779	732	727
base	815.2667	1230	1091	758	562	611	697	650	421	916	1068	1044	757	864	927	633
base	730.2	1063	637	506	1019	798	529	661	552	803	745	622	900	629	834	655
base	767	828	779	720	557	656	523	1811	546	588	681	880	665	661	798	812
base	643	752	728	622	711	578	588	573	466	550	0	549	668	723	1138	999
base	781.2667	912	773	795	630	726	598	964	435	732	795	1160	889	791	776	743
base	835.2667	952	691	743	724	574	628	1585	469	1290	875	828	750	754	777	889
base	960.1333	1049	1553	1034	1031	700	713	1500	684	594	1135	1243	840	517	900	909
base	819.8667	767	923	947	761	928	645	975	479	1561	736	638	687	700	779	772
mean	768.1846															
sem	716.4667	856	642	657	781	557	664	884	521	1059	776	620	713	672	877	468
sem	740.0667	909	548	936	637	645	800	824	524	650	880	626	883	538	692	1009
sem	726.8	1075	686	561	690	773	512	839	560	844	1077	551	610	640	707	777
sem	670.6667	1046	661	581	763	486	450	767	724	978	598	496	749	513	694	554
sem	747.2667	1134	547	629	791	758	577	1037	521	672	1068	661	956	609	677	572
sem	799.0667	737	1183	598	954	1008	753	690	485	1418	669	689	647	549	933	673
sem	769.1333	874	655	738	619	521	546	1860	550	653	1125	717	702	542	777	658
sem	708.8	992	824	773	694	696	629	665	479	853	672	548	683	908	744	472
sem	695.9333	903	668	702	563	646	802	675	509	636	972	644	713	536	905	565
sem	827.4	940	935	1438	807	978	794	908	623	765	869	724	827	550	632	621
sem	710.9333	964	703	846	613	535	610	624	495	1146	684	525	706	471	993	749
sem	935.4667	882	1063	675	762	632	882	1428	755	1272	1350	1169	1065	506	880	711
sem	720.2667	1083	658	648	775	524	1129	1042	853	0	868	671	601	530	735	687

Towards Arabic Named Entity Recognition Tool

Nouran Khallaf [*1], Sameh Alansary[*2]
** Nouran - Sameh Phonetics Department, Faculty of Arts, Alexandria University*
El-Shatby, Alexandria, Egypt
[1]*Nouran_khallaf@yahoo.com*
[2]sameh.alansary@bibalex.org

***Abstract*— This paper investigates the area of Named entity recognition(NER) which have a magnificent effect in the improvement of the performance of many natural language processing applications. The NER task in a specific language is achieved by collecting information about the language and its linguistic characteristics. This paper takes interest in describing this task in general giving much attention on the efforts applied to Arabic language in specific. It therefore discusses the complexity of Arabic NER and the challenges that faces this task along with the available language resources. It also discusses the available Arabic tools and systems. In addition to shedding some light on an ongoing research that aims at building an Arabic NER system. Adopting the Rule based approach, dictionary based approach, classification based approach and sequence based approach. Each of them is applied for a specific Named entity types as each type could be recognized through deferent approach.**

Keywords: Name entity, Name entity recognition, Arabic name entity tag set**.**

1 INTRODUCTION

Corpus Annotation originally means attaching linguistic information on lexical level (part of speech tagging), grammatical functions or non-linguistic information. Annotation of named entities is considered as a part of POS-tagging as a lexical element is POS-tagged as NE if they belong to one of the categories of proper nouns. A Named Entity (NE) is a word, or sequence of words that can be classified as a name of a person, organization, location, date, time, percentage or quantity(Solorio T, 2011).Named entities can be valuable in several natural language Processing (NLP) such as Information Retrieval and Question Answering tasks text clustering Named entity recognition (NER) systems aim to automatically identify and classify the proper nouns in text. The Named Entity Recognition (NER) task has been gaining huge attention in Natural Language Processing (NLP) as it is proved to have a magnificent improvement in the performance of many natural language processing applications.

Away from the Natural Language Processing application and tools, one of the most important fields which provides numerous improvement in such applications is Named Entity Recognition (NER). This field is my concern in the paper. The first research papers in this field was presented by Lisa F. Rau (1991) at the Seventh 2 IEEE Conference on Artificial Intelligence Applications. Rau's paper describes a system to "extract and recognize company names"[1]. Then linguists noticed that it is important to recognize such units like names of person, organization and location, and numeric expressions including time, date, money and percent expressions. The term Named Entity (NE), first introduced in 1995 by the Message Understanding Conference (MUC-6), is widely used in the field of Natural Language Processing and Information Retrieval. Since 1995, a lot of studies have addressed NE recognition, tagging and classification.

Named entity recognition (NER) systems aim to automatically identify and classify the proper nouns in text. Named Entity Recognition (NER) is the task of detecting and classifying proper names within texts into predefined types, such as Person, Location and Organization names [2]. NER systems play a significant role in many areas of Natural Language Processing (NLP) such as question answering systems, text summarization and information retrieval.

This paper reports research into the Arabic Named Entity Recognition systems. Challenges of named entity recognition, Approaches and algorithms for NER have been analyzed and compared on the theoretical level, and resources, methods and tools for the practical evaluation and comparison of Arabic NER have been designed and implemented. Reporting the available tools that could be used for Arabic NE tagging and the resource that could be used in building such a tool. Finally presenting the proposed system architecture a way to build an Arabic NER tool simple, fast and accurate.

2 NAMED ENTITY RECOGNITION CHALLENGES

Although Named entity recognition seems to be a simple task, faces a number of challenges. The NER task in a specific language is achieved by collecting information about the language and its linguistic characteristics. To understand the complexity of Arabic NER we need to classify the challenges into two challenges as follows:

A. Name entity challenges

1) Variation of NEs (Detect the boundaries) Example: محمد على ,الاستاذ محمد على ,محمد.[Mohamed, Mr. Mohamed Ali, Mohamed Ali]
2) Ambiguity of NE types for example Locations and person names can be the same (Metonyms) refer to it as Referential Relativity Interpretation [3] – Example:شارع محمد على ,الاستاذ محمد على. [Mr. Mohamed Ali, Mohamed Ali St.]
3) Synonymy arises when different names refer to the same entity (Denominational Stability) Example: America and U.S referring to the United States of America
4) Issues of style, structure, domain, and genre.
5) Punctuation, spelling, spacing, formatting.

To conclude, to be able to identify NEs successfully, there would be a need to analyze the context not only the NEs but also the surrounding lexical items with their Part of Speech categories. As the most problematic classified words in POS tagger is the proper nouns.

B. Arabic Named entity recognition challenges

Arabic Language is one of the Semitic languages. It is the mother tongue of 317 million Arabs and the religious language of more than 1 billion Muslims.[4]. Semitic family is part of the Afro-Asiatic family and its first written form was introduced in the third millennium BC. The most widely spoken Semitic language today is Arabic It is written from right to left. It has 28 letters phonemes each letter might has three different shapes according to its place in a word (initial-middle-final), three long vowels and five main short vowels represented by diacritic marks placed above or below the letters. Only three letters are not affected by that feature. Six letters in the alphabet have only two possible forms because only preceding letters could connect to them; these six letters cannot be connected to the following letters. Arabic NER faces major challenges [5] :

1) No capitalization: Absence of capital letters in the orthographic form of Arabic, unlike English that the presence of capital letters eases the process of detecting the Named Entities.

2) Morphological patterns: Challenges come from Arabic morphological patterns because Arabic is highly inflectional language; often a single word has more than one affix such that it may be expressed as a combination of prefix(s), lemma, and suffix(s) thus a huge training corpus is required in order to obtain a high accuracy (See Fig.1) which shows how a whole English sentence could be represented by only one Arabic word. The prefixes are articles, prepositions, or conjunctions. In order to analyze data there are two methods: (1) Light stemming: in which all prefixes and suffices are deleted to reach the stem and know the meaning of it. In this solution treating the affixes as stop words does not affect the meaning. (2) Word segmentation: consists of separating the different components of a word by a space character.

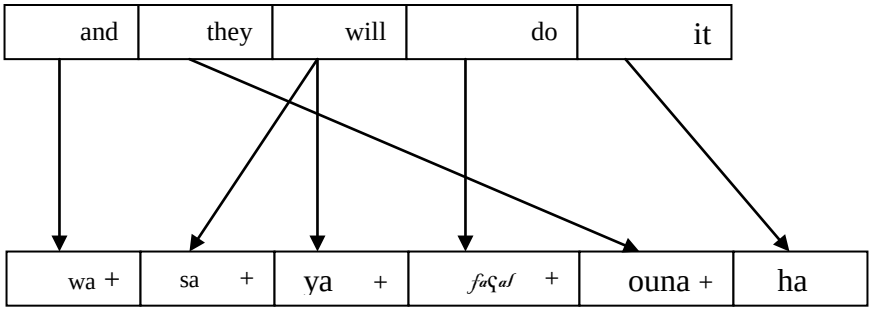


Figure1 representing a whole English sentence " and the will do it" with one Arabic word "wasayafa3alounaha"

3) Spelling variations: Arabic character may have up to three different forms; each form corresponds to a position of the character in the word (beginning, middle or end of the word). Arabic has some variants in spelling and orthographic forms. For example, "جرام-غرام" "Gram" is a spelling deviation and "استراليا-أستراليا" "Australia" is an orthographic variant. And using The Kashida [ـ] is a special character for lengthening a letter.

4) *Ambiguity:* Arabic texts have different sorts of ambiguities (different meanings) according to the diacriatization of the word. For example, "ذهب"/"Zahab" in Arabic may be used as a verb goes, or noun gold.
5) *Word Order:* Arabic is a free-word order language, in which many named entities cannot be recognized to have a specific pattern could be detected through. Sometimes the proper noun in the phrase appears next to the keyword or appears after four or five words after the keyword or appears before the keyword or completely omitted from the phrase.[6]. For example,"الرئيس السابق جورج بوش" "جورج بوش الرئيس السابق". [Former President George W. Bush ,George W. Bush, former President]
6) *Transliteration of loan and borrowed words:* Arabic has a number of loan Latin words. The difference between Loan words and borrowed words is that the loan words are words from another language representing a new concept. On the other hand, the borrowed words are words that users use instead of their own language's vocabulary. For example, "ألاسكا باى"or "خليج ألاسكا" [Alascapay]the first transliteration would be difficult to detect.

3 PREVIOUS WORKS (See Table 1)

Since 1996 there was acceleration in NER systems through many conferences in many languages such as German, Japanese, Greek, French, Italian, Basque, Bulgarian, Catalan, Cebuano, Danish, Hindi, Korean, Romanian, Russian, Swedish, Turkish, Portuguese, and Arabic. Arabic started to receive attention since 2005[7].Most of the work in NER has concentrated on limited domains and textual genres such as news articles and web pages [8]. The following survey is structured through three main algorithms (Rule-based, ML-based and Hybrid approach). Here they are chronologically classified started from 1998 to 2012.

TABLE 1
SUMMARY OF PREVIOUS WORKS IN ANER (ARABIC NAMED ENTITY RECOGNITION)

Participant	Used Corpora		NE types	Evaluation		
	Tagging	Evaluation		P.	R.	F-m
Arabic Rule-based NER						
(Maloney and Niv 1998)[9]	3214 tokens		Person, Organization, Location, Number and Time.	73.0%	93.5%	82.0%
(Samy, Moreno et al. 2005)[10]	900 sentence (F/A)pair	300 sentence (F/A)pair	Proper Names ,Toponyms , Acronyms , Jobs , Organizations, Dates	84%	97.5%	89.3%
(Nezda, Hickl et al. 2006)[11]	800,000 word	600,000 words	18 NEs drived from Numeric Expression, Temporal Expression, Quantities, Names, Artifacts	——	——	85%
(Mesfar 2007)[12]	"Le Monde Diplomatique" corpus		Person, Location, Organization, Currency, Temporal expression	——	——	87%
(Shaalan and Raza 2007)[13]	ACE , Treebank corpus 472617 entries		Person Named Entity	86.7%	89%	87.8%
(Shaalan and Raza 2008, Shaalan and Raza 2009)[14, 15]	300,000 words	397,069 words	location, company, date, time, price, measurements, phone number , ISBN and file name.	91.68%	93.53%	92.26%
(Traboulsi 2009)[16]	245,213,037 words		——	——	——	——
(O'Steen and Breeden 2009)[17]	ANERcorp (150,285 tokens) two equal sets		Person, Location, Organization and Micsistones	76.4%	33.1%	45.9%
(Al-Shalabi, Kanaan et al. 2009)[18]	——	20 articles AlRaya newspaper	Location, Person, Event, Organization, Temporal, Equipment and Scientific	accuracy was 86.1%		
(Elsebai, Meziane et al. 2009)[6]	——	700 news articles from Aljazeera	Person Named Entity	——	——	89%
(Alkharashi 2009)[19]	80,000 names and surnames of Saudi	20,000 names and surnames of Saudi	Person Named Entity	——	——	——
(Zaghouani, Pouliquen et al. 2010)[20]	One million word	34,000 tokens	Person, Location , Organizations, Dates , Numeric expression	87.17%	65.74%	74.95%

(Hamadou, Piton et al. 2010)[21]	——	hundred sports texts	names of athletic venues: stadiums, arenas, pools, tracks…etc	97%	95%	96%
(Fehri, Haddar et al. 2011)[22]	——	4000 texts of sport	Player Names, Team Names, Sport Names	98%	90%	94%
		300 texts of education	——	98%	70%	82%
(Asharef, Omar et al. 2012)[23]	13300 words	6500 words	person names, organization names, location names and time	91%	89%	89.46%
(Shihadeh and ünter Neumann 2012)[24]	ANERcorp corpus		persons, locations and organizations	43.59%	16.63%	33.77
(Algahtani 2012)[4]	60,000 words	10.000 words	persons, locations and organizations	83.40%	70.06%	76.39%
			Person Named Entity	81.81%	70.24%	75.59%
Arabic ML-based NER						
(Zitouni, Sorensen et al. 2005)[25]	ACE 2003, ACE 2004		——	accuracy was 69.2%		
(Benajiba, Rosso et al. 2007) [26]	125,000 ANERcorp	25,000 ANERcorp	persons, locations and organizations	63.21%	49.04%	55.23%
(Benajiba and Rosso 2007) [27]	125,000 ANERcorp	25,000 ANERcorp	persons, locations and organizations	70.24%	62.08%	65.91%
(Benajiba and Rosso 2008)[28]	125,000 ANERcorp	25,000 ANERcorp	Person, Location, Organization and Micsistones	86.90%	72.77%	79.21%
(Benajiba, Diab et al. 2008)[29]	ACE 2003, 2004 ,2005 corpora and UPV-corpus		Person, Location, Organization and Micsistones	——	——	83.5%
(**Benajiba, Diab et al. 2009)**	ACE 2003, 2004 ,2005 corpora and UPV-corpus		Person, Location, Organization and Micsistones	——	——	82.17%
(Benajiba, Zitouni et al. 2010)[30]	LDC 941,282		Person, Location, Organization and Micsistones	——	——	84.32%
(Abdul-Hamid and Darwish 2010)[31]	ANERcorp , ACE 2005		persons, locations and organizations	86%	69%	76%
(**Koulali and Meziane 2012)**[32]	Arabic unvowelized documents		CoNNL2003 tag set	92.54%	77.07%	83.20%
(Alotaibi and Lee 2012)[33]	ANERcorp	25674 tokens	two binary tags 'U-NE' for each named entity and 'O' for otherwise	96.06%	82.25%	88.62%
(Mohammed and Omar 2012)[34]	ANERcorp		Person, location, company, date, time, price, measurements, phone number , ISBN and file name.	89.93%	34.25%	92.36%
Arabic Hybrid approach						
(Abuleil 2006)[35]	——	200 articles from the Al-Quds newspaper	persons, locations and organizations	accuracy was 98.6%		
(**Noha Ahmed, Ali Farghaly et al. 2011)[36]**	ANERcorp , ACE 2003		persons, locations and organizations	90.5%	87.05%	88.77%
(Oudah and Shaalan 2012)[37]	ACE, ATB v2.0, ANERcorp		person, location, organization, date, time, price, measure, percent, phone number, file name, ISBN	——	——	90.9%

4 AVAILABLE TOOLS

A. BBN's IdentiFinder Text Suite™ (See Fig. 2)

Tool that analyzes electronically-stored text to locate names of corporations, organizations, people, and places, including variations in names. Using statistical methods after the failing of the rule-based system. The system recognize 26 types of named entity with detection and classification for English, Arabic and Chinese languages. could be accessed through [http://bbn.com/technology/speech/identifinder]

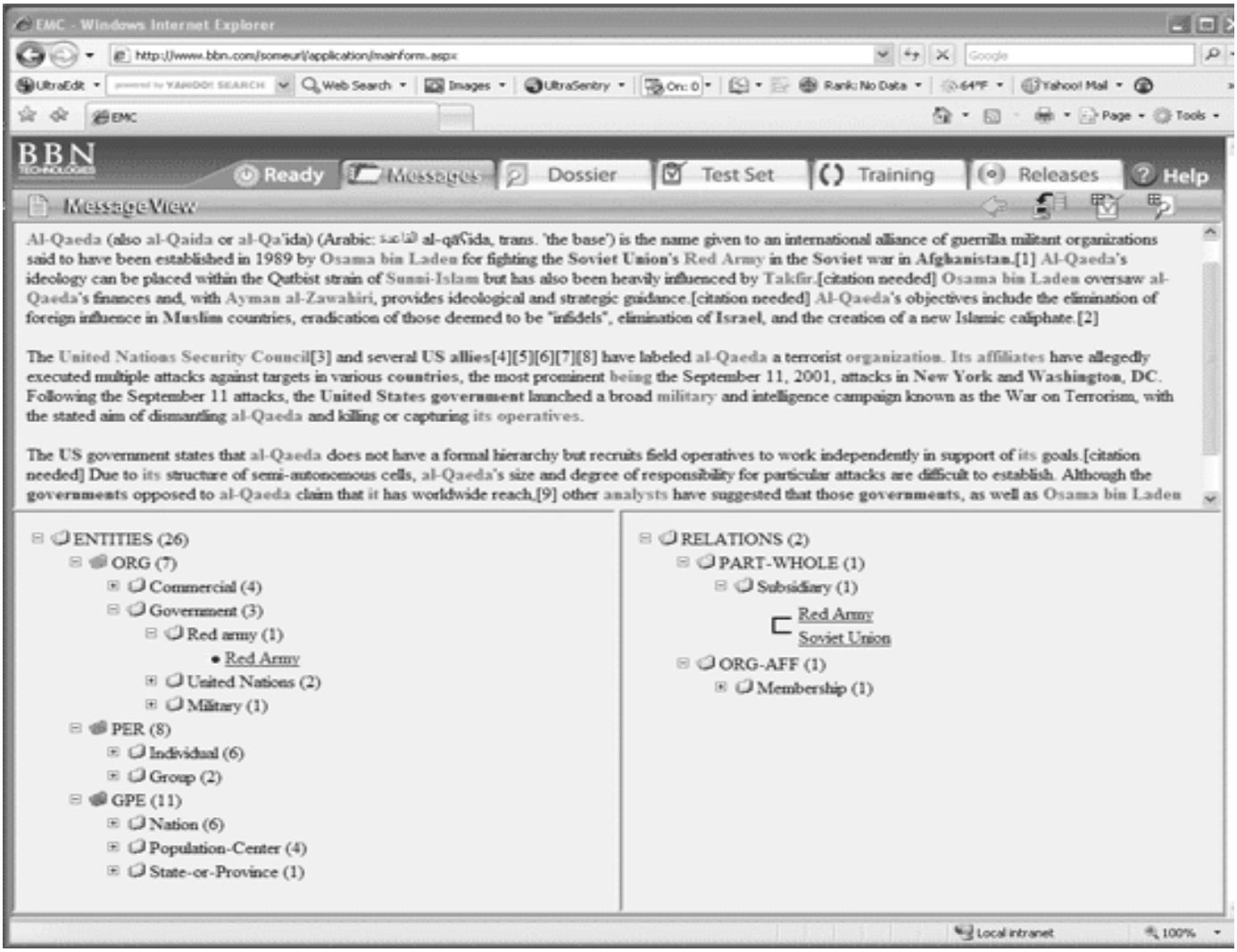


Figure 2 Screenshot of an application developed with BBN Identifinder Text Suite™

B. NetOwelExtractor (NetOwel)

NetOwl Extractor tag entities (over 100 types of entities) and its relationships in unstructured text. It supports multiple *domains* (Business, Compliance, Customer Experience Management, Cyber Security, Finance, Homeland Security, Intelligence, Law Enforcement, Life Sciences, Litigation Support, Military, National Security, Opinion Mining, Politics and Sentiment Analysis)and *languages* (English, Arabic, Chinese, French, Korean, Persian (Farsi, Dari), Russian and Spanish) . It offers English translation of entities extracted from foreign language. [http://www.sra.com/netowl/entity-extraction/]

C. Rosette Entity Extractor (REX)

This named entity recognition software provides semantic tagging to find entities in text using a hybrid approach of statistical methods, regular expressions and gazetteers. it supports 16 languages including Arabic. Capable to classify 18 named entities for 6 domains (health, medical, life sciences, financial, and manufacturing).(See Fig. 3)

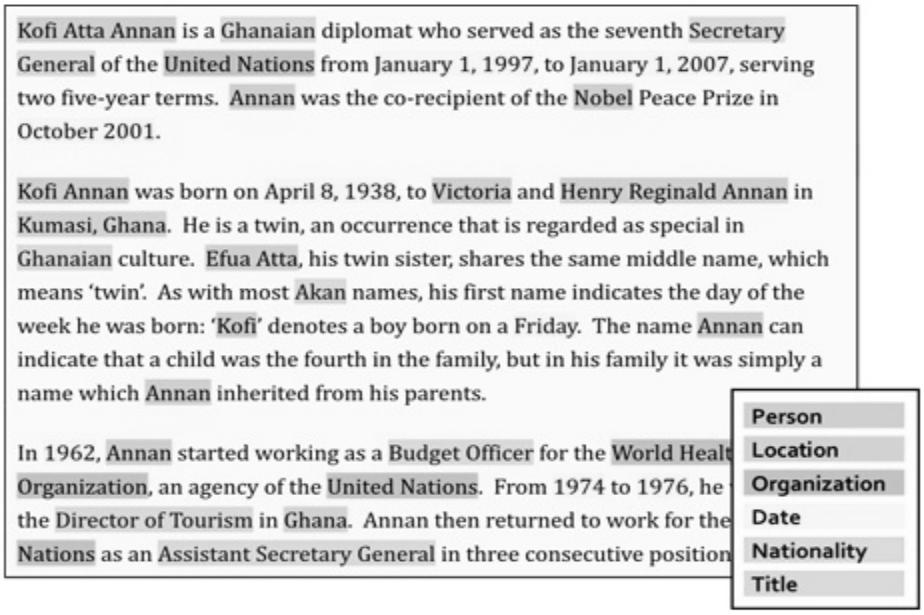


Figure3 Text tagged sample of REX

D. Annoqt

Semi- automatic annotation tool for Arabic, English and French. Annoqt is natively multilingual. It particularly handles gracefully right-to-left languages and non-latin scripts. Supports overlapping entities (or even entities completely embedded inside other entities). Has been developed by the CEA (which is a French government-funded technological research organization.).[38] (See Fig. 4)

E. CICEROARABIC

It is the first wide coverage named entity recognition (NER) system for Modern Standard Arabic. Capable of classifying 18 different named entity classes with over 85% F. CICEROARABIC utilizes a new 800,000- word annotated Arabic newswire corpus in order to achieve high performance without the need for hand-crafted rules or morphological information.[11].

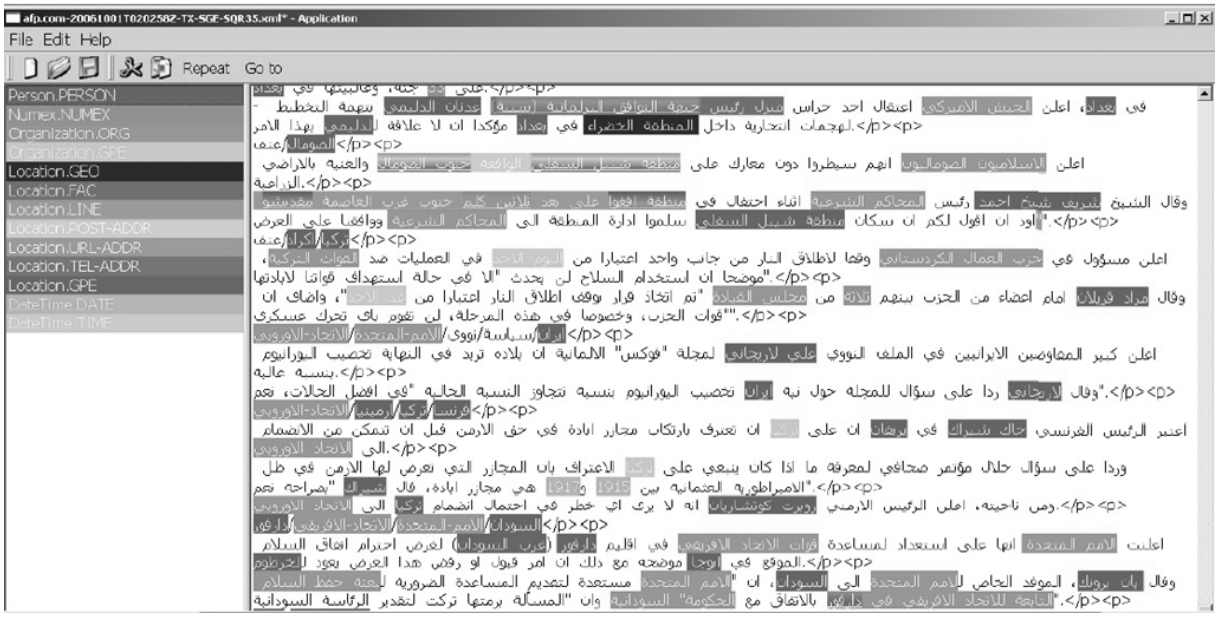


Figure 4 Screenshot of Annoqt tool

F. ANEE: Arabic Named Entity Extraction

Extracts critical information from large amounts of structured and unstructured data using human semantic concepts using a hybrid approach. provides tagging with 25 main NE categories and 100 Subcategories and setting the relationships between entities. It is available as a system development kit (SDK) for integration into an existing application, regardless of platform, ANEE also can be used as a stand-alone application. (See Fig. 5)

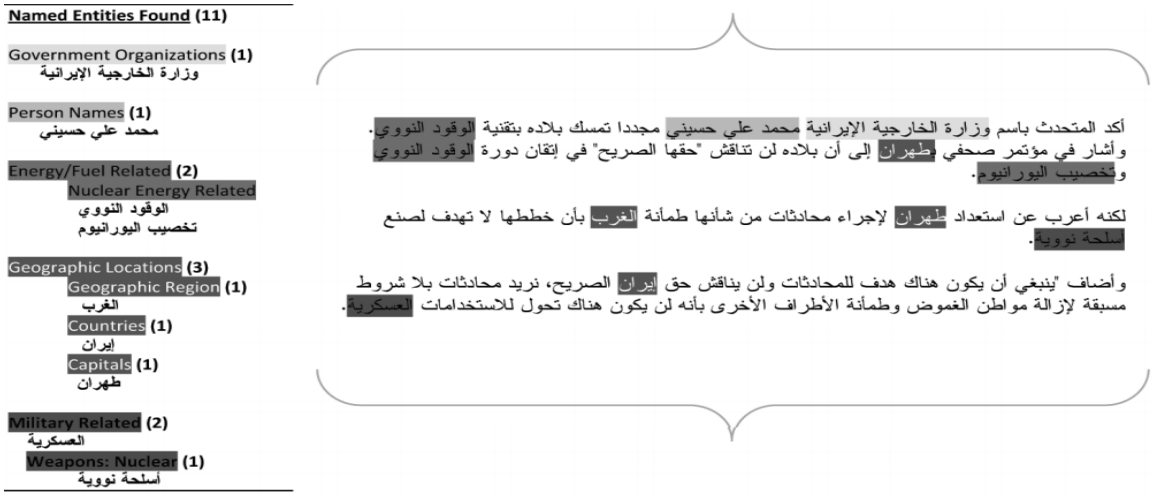


Figure 5 A Sample Entity Extraction Output of ANEE

5 LANGUAGE RESOURCES RELEVANT TO ARABIC NAMED ENTITY RECOGNITION

A. Annotated Corpora for NER purposes :

Some of the Corpora were annotated for the purpose of Arabic NER :(See Table2)

TABLE 2
ANNOTATED CORPORA WITH NES

Corpus	Developer	NEs Types	Number of tokens	Languages
KALIMAT 1.0	Dr Mahmoud El-Haj	person, location, rganization and miscellaneous	20,291	Arabic
ANERcorp	Yassine Benajiba	person, location, organization and miscellaneous	150,286	Arabic
(Mostefa, Laïb et al. 2009)[38]	French Ministry of Research	5 higher classes and 11 subclasses.	1,462,085	Arabic, English, French
ACE Data	ACE 2003-2005	Person, Organization, Facility and Geo Political Entity	297366	Arabic
OntoNotes 4.0[39]	GALE (Global Autonomous Language Exploitation) program	(Person, NORP[1], facility, Organization, GPE, Location, Product, Event, Art, law and Language)	400k of Arabic	English, Chinese, and Arabic
ALTEC	Altec Organization	God, Organization, Location, Facility, Product, Event, Natural object, Disease, Color, Jobs, Nationality, Timex , Numex	5,000,000	Arabic
Arabic WordNet	Unknown	Unknown	23,481	Arabic

B. Gazetteers:

Some of the available Gazetteers that could be used in building such tool (See Table 3)

TABLE 3
NE AVAILABLE GAZETTEERS

Gazetteer	Number of entities
ANERgazet	Location Gazetteer: 1,950
	Person Gazetteer : 1,920
	Organization Gazetteer: 262
(Attia, Toral et al. 2010)[40]	45,000 Arabic NEs.
HeiNER	1.5 million NEs in 16 languages
JRC-Names[41]	205,000 person and organization names + 205,0000 of spelling variants
Wentland et al. (2008)	1.5 million English names linked to anther fifteen languages
Geonames.de	Countries and Languages of the World
CJKI's Database of Arab Names (DAN) v3.0	Seven million Arab Names in Arabic

C. Toolkits:

*1) Natural Language Toolkit -NLTK: NLTK was originally created in 2001 as part of a computational linguistics course in the Department of Computer and Information Science at the University of Pennsylvania. Components of NLTK , **Corpora** : more than 300Mb annotated data sets widely used in natural language processing contains (*

Brown Corpus, Carnegie Mellon Pronouncing Dictionary, CoNLL 2000 Chunking Corpus, Project Gutenberg Selections, NIST 1999 Information Extraction: Entity Recognition Corpus, US Presidential Inaugural Address Corpus, Indian Language POS-Tagged Corpus, Floresta Portuguese Treebank, Prepositional Phrase Attachment Corpus, SENSEVAL 2 Corpus, Sinica Treebank Corpus Sample, Universal Declaration of Human Rights Corpus, Stopwords Corpus, TIMIT Corpus Sample. Treebank Corpus Sample ...etc) all of those corpora could be accessed through the second component of NLTK, ***Code*** *: 50k lines of code of corpus readers, tokenizers, stemmers, taggers, chunkers, parsers, wordnet, ... etc.*

2) *General Architecture for Text Engineering GATE: A full-lifecycle open source solution for text processing. This a language engineering environment developed at the University of Sheffield. Its first release in 1996. There are set of resources in GATE:* ***ANNIE*** *(A Nearly-New IE system) which consists of main processing resources for information extraction such as : tokenize, sentence splitter , POS tagger ,gazetteer, semantic tagger, document reset and finite transducer,* ***ANNIE Gazetteer*** *allows the automatic annotation of place with list of places from 18th century reports from the Old Bailey in London,* ***JAPE*** *(Java Annotation Pattern Engine) Regular expressions over annotations with Finite state transduction over annotations based on regular expressions,* ***ANNIC:*** *ANNotations-In-Context a full-featured annotation indexing and retrieval system,* ***AnnotationDiff*** *tool used for evaluation for annotation by comparing two annotated documents with generating figures for Precision, Recall, F-Measure and false positives, and* ***Benchmarhing*** *tool also for evaluation but it enables evaluation to be done over a whole corpus rather than a single document its output is written to an HTML file in a tabular form. Balance Distance Measure (BDM) Ontology Tool*

6 PROPOSED SYSTEM

The data used in building the system was a part of ALTEC corpus (3million word) that conducts 10 fields of divided into 275,000 for training and 75,000 for testing. Using the tag set proposed by ALTEC organization to annotate ALTEC corpus. The tag set built for the sake of Arabic, in light of the BBN tag set and Sekine's extended hierarchy. It is a hierarchy tag set of three levels in which the first level consists of 15 types and the second level consist of 39 and the third level consist of 72 sub-types to conduct a 104 tags. Tagged through SGML tags but with specific tag names built according to the adapted tag set. With using the tag "Other" to present a flexible tag set.

The proposed system consists of four main processes. First, preprocessing which conduct the segmentation rules and normalization. Second, analysis that consists of adding morphological features and noun phrase chucker. Third, the NE tagging with Rules and gazetteers. Finally, the filtration process (See Fig. 6). In the normalization preprocessing phase all the variants of one letter are normalized to only one shape for example (إ،أ،ا،آ) becomes (ا).

Some Issues in tagging process:

The most important thing that we should concentrate on while tagging NEs is that each type should be treated separately and each type has its specific way of tagging which means

1) Some NE types should be tagged before others for example: the type [Event-Occasion-Sport] before [Location-Region-Continental] as the second may be a part of the first such as{ بطولة دورى القارة الافريقية ,The African continent League Cup} in which {القارة الافريقية, The African continent } is a part of the main NE.
2) Some of specific NEs are listed in the gazetteers such as{ افريقيا , Africa} should be listed but {الافريقية, The African} should not be listed but it should be detected through the trigger word (القارة , continent) because the NE افريقيا could come alone without a trigger word but NE الافريقية could not come without its trigger word to be tagged with the type [Location-Region-Continental]
3) Some of trigger words are preceding or following the NEs.

4) Person Names should be treated in a different way considering the all type of person names Arabic person names given names(ism), relative adjective (Nisba), epithenton (laqab), teknonym (names that have been derived from the child's given names), patronymics names (it is derived from the father's name).
5) Most of non Arabic words are not analyzed with the morphological analyzer so most of non analyzed words should be translated first to know its ne type.
6) The main architecture could be modified depending on the NE type and its structure.

The filtration process, is an important process to make sure of the tagging had been done and detect other untagged words. For Example, searching with tagged words that have trigger word to detect the same words in non-tagged words. Although this searching would be good for "Location" detection, it could not be done with the tag type "Titles" as it could be a word within the text.

A. Evaluation of the system:

The named entity is correct only if it is an exact match of the corresponding entity in the solution ignoring boundaries[42].Based on micro-averaged F-Measure with:

1. Precision: the percentage of named entities found by the system that are correct.
2. Recall: the percentage of named entities present in the solution that are found by the system.
3. Micro-averaged F-Measure.

Using ***AnnotationDiff*** tool which is a part of **General Architecture for Text Engineering GATE [**is freely available for download from http://gate.ac.uk**]** used for evaluation for annotation by comparing two annotated documents with generating figures for Precision, Recall, F-Measure and false positives.

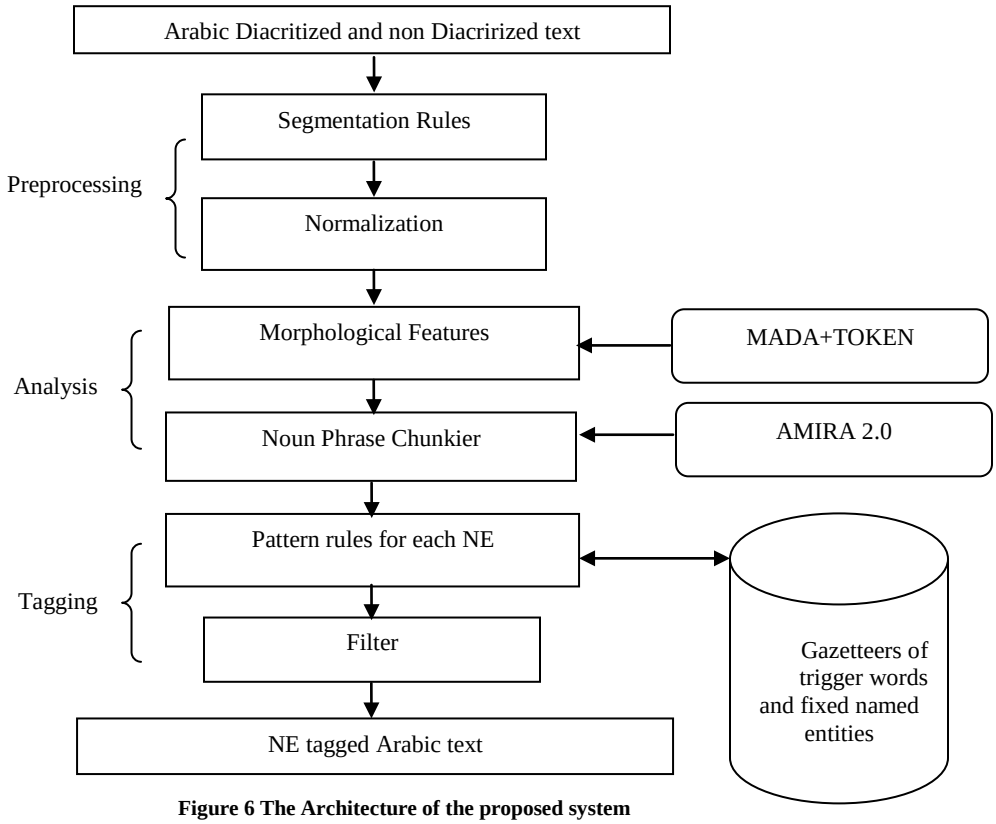


Figure 6 The Architecture of the proposed system

REFERENCES

[1] Rau, L.F. *Extracting company names from text*. in *Artificial Intelligence Applications, 1991. Proceedings., Seventh IEEE Conference on*. 1991. IEEE.

[2] Nadeau, D., *Semi-supervised named entity recognition*. 2007, Citeseer.

[3] Fort, K., M. Ehrmann, and A. Nazarenko. *Towards a methodology for named entities annotation*. in *Proceedings of the Third Linguistic Annotation Workshop*. 2009. Association for Computational Linguistics.

[4] Algahtani, S.M., *Arabic named entity recognition: a corpus-based study*. Doctoral dissertation,, University of Manchester, 2012.

[5] Riaz, K. *Rule-based named entity recognition in Urdu*. in *Proceedings of the 2010 Named Entities Workshop*. 2010. Association for Computational Linguistics.

[6] Elsebai, A., F. Meziane, and F.Z. Belkredim, *A rule based persons names Arabic extraction system.* Communications of the IBIMA, 2009. **11**(6): p. 53-59.

[7] Nadeau, D. and S. Sekine, *A survey of named entity recognition and classification.* Lingvisticae Investigationes, 2007. **30**(1): p. 3-26.

[8] Samir Abdelrahman, M.H., Marwa Magdy and Aly Fahmy, *Information Extraction*, in *The Pre-Swot Analysis*. 2012, ALTEC-Organization: Cairo, Egypt.

[9] Maloney, J. and M. Niv. *TAGARAB: a fast, accurate Arabic name recognizer using high-precision morphological analysis*. in *Proceedings of the Workshop on Computational Approaches to Semitic Languages*. 1998. Association for Computational Linguistics.

[10] Samy, D., A. Moreno, and J.M. Guirao. *A proposal for an Arabic named entity tagger leveraging a parallel corpus.* in *International Conference RANLP, Borovets, Bulgaria*. 2005.

[11] Nezda, L., et al., *What in the world is a Shahab? Wide coverage named entity recognition for Arabic.* Proccedings of LREC, 2006: p. 41-46.

[12] Mesfar, S., *Named entity recognition for arabic using syntactic grammars*, in *Natural Language Processing and Information Systems*. 2007, Springer. p. 305-316.

[13] Shaalan, K. and H. Raza. *Person name entity recognition for arabic*. in *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*. 2007. Association for Computational Linguistics.

[14] Shaalan, K. and H. Raza, *Arabic named entity recognition from diverse text types*, in *Advances in Natural Language Processing*. 2008, Springer. p. 440-451.

[15] Shaalan, K. and H. Raza, *NERA: Named entity recognition for arabic.* Journal of the American Society for Information Science and Technology, 2009. **60**(8): p. 1652-1663.

[16] Traboulsi, H. *Arabic named entity extraction: A local grammar-based approach*. in *Computer Science and Information Technology, 2009. IMCSIT'09. International Multiconference on*. 2009. IEEE.

[17] O'Steen, D. and D. Breeden, *Named Entity Recognition in Arabic: A Combined Approach*. 2009.

[18] Hamish, C., et al. *GATE: an Architecture for Development of Robust HTL applications*. in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2002.

[19] Alkharashi, I. *Person Named Entity Generation and Recognition for Arabic Language*. in *Proceedings of the Second International Conference on Arabic Language Resources and Tools. Cairo, Egypt*. 2009.

[20] Zaghouani, W., et al. *Adapting a resource-light highly multilingual Named Entity Recognition system to Arabic*. in *LREC*. 2010.

[21] Hamadou, A.B., O. Piton, and H. Fehri, *Recognition and translation Arabic-French of Named Entities: case of the Sport places.* arXiv preprint arXiv:1002.0481, 2010.

[22] Fehri, H., K. Haddar, and A. Ben Hamadou. *Recognition and translation of Arabic named entities with NooJ using a new representation model*. in *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*. 2011. Association for Computational Linguistics.

[23] ASHAREF, M., et al., *ARABIC NAMED ENTITY RECOGNITION IN CRIME DOCUMENTS.* Journal of Theoretical and Applied Information Technology, 2012. **44**(1).

[24] Shihadeh, C. and G. ünter Neumann. *ARNE-A tool for Namend Entity Recognition from Arabic Text*. in *The Fourth Workshop on Computational Approaches to Arabic Script-based Languages*. 2012.

[25] Zitouni, I., et al. *The impact of morphological stemming on Arabic mention detection and coreference resolution*. in *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*. 2005. Association for Computational Linguistics.

[26] Benajiba, Y., P. Rosso, and J.M. Benedíruiz, *Anersys: An arabic named entity recognition system based on maximum entropy*, in *Computational Linguistics and Intelligent Text Processing*. 2007, Springer. p. 143-153.

[27] Benajiba, Y. and P. Rosso. *ANERsys 2.0: Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information*. in *IICAI*. 2007.
[28] Benajiba, Y. and P. Rosso. *Arabic named entity recognition using conditional random fields*. in *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*. 2008.
[29] Benajiba, Y., M. Diab, and P. Rosso. *Arabic named entity recognition: An svm-based approach*. in *Proceedings of 2008 Arab International Conference on Information Technology (ACIT)*. 2008.
[30] Benajiba, Y., et al. *Arabic named entity recognition: using features extracted from noisy data*. in *Proceedings of the ACL 2010 conference short papers*. 2010. Association for Computational Linguistics.
[31] Abdul-Hamid, A. and K. Darwish. *Simplified feature set for Arabic named entity recognition*. in *Proceedings of the 2010 Named Entities Workshop*. 2010. Association for Computational Linguistics.
[32] Koulali, R. and A. Meziane, *A combined Approach to Arabic Named Entity recognition Using SVM and Pattern Extracted method applied to Topic Detection.* 2012.
[33] Alotaibi, F. and M. Lee. *Using Wikipedia as a resource for Arabic named entity recognition*. in *Rabat, Morocco. In Proceeding of the 4th International Conference on Arabic Language Processing (CITALA12)*. 2012.
[34] Mohammed, N.F. and N. Omar, *Arabic Named Entity Recognition Using Artificial Neural Network.* Journal of Computer Science, 2012. **8**(8): p. 1285.
[35] Hamish, C., et al. *GATE: an Architecture for Development of Robust HLT Applications*. in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2002.
[36] Noha Ahmed, Ali Farghaly, and A. Fahmy, *A survay of Named Entities heirarichy*, in *Arabic Language Technology International Conference (ALTIC) Program*. 2011, ALTEC. p. 311-322.
[37] Oudah, M. and K.F. Shaalan. *A Pipeline Arabic Named Entity Recognition Using a Hybrid Approach*. in *COLING*. 2012.
[38] Mostafa, D., et al., *A multilingual named entity corpus for Arabic, English and French.* MEDAR, 2009. **2009**: p. 2nd.
[39] Weischedel, R., et al., *OntoNotes Release 4.0*. 2010, Tech. rept. BBN Technologies.
[40] Attia, M., et al., *An automatically built named entity lexicon for Arabic.* 2010.
[41] Steinberger, R., et al. *JRC-NAMES: A Freely Available, Highly Multilingual Named Entity Resource*. in *RANLP*. 2011.
[42] Nothman, J., *Learning named entity recognition from Wikipedia*. 2008, The University of Sydney Australia 7.

نحو التعرف الآلى على الكيانات الاسمية فى اللغة العربية

نوران خلاف[١] ، سامح الانصارى[٢]
قسم الصوتيات واللسانيات، كلية الآداب، جامعة الاسكندرية
الشاطبى، الاسكندرية، مصر
١ nouran_khallaf@yahoo.com
٢ sameh.alansary@bibalex.org

ملخص:
هذه الورقة البحثية تقدم بحث في التعرف الآلي للكيانات الاسمية فى اللغة العربية (NER) والتي لها تأثير رائع في تحسين أداء العديد من تطبيقات معالجة اللغة الطبيعية. يتم التعرف على الكيانات الاسمية في لغة معينة من خلال جمع المعلومات عن اللغة وخصائصها اللغوية. تهتم هذه الورقة بوصف عملية التعرف الآلى في العام مع إعطاء الاهتمام الاكبرللجهود التى تمت على اللغة العربية بشكل خاص. وبالتالي فإنها تناقش تعقيد التعرف الآلي للكيانات الاسمية فى اللغة العربية والتحديات التي تواجه هذه المهمة جنبا إلى جنب مع الموارد اللغوية المتاحة للغة العربية. كما يناقش الأدوات والنظم العربية المتاحة. بالإضافة إلى تسليط بعض الضوء على البحث الجاري و اللذي يهدف إلى بناء نظام التعرف الآلي للكيانات الاسمية فى اللغة العربية. اعتماداً على النهج القائم على قواعد اللغوية، والقائم على القواميس

اللغوية ، والقائم على التصنيف. حيث سيتم تطبيق كل واحد منهم لمجموعة أنواع محددة للكيانات الاسمية على أساس أفضل طريقة للكشف عن كل نوع على حدة.

Performance of Different Speech Coders over WiMAX and LTE

Neamat. A. Kader[1] , Nermeen. A. Rasmy[2], Heba. Mourad[3]

[*]*Electronic and Communication Department, Faculty of Engineering, Cairo university*
Giza, Egypt
[1]nemat2000@hotmail.com
[2]fantoum2003@yahoo.com
[3]hmourad.mourad@gmail.com

***Abstract*-This paper compares the performance of various voice coders and evaluates the quality of these coders over_WiMAX and LTE. Different codecs namely ITU-T G.711, ITU-T G.729, ITU-T G.723.1 and iLBC are used in the simulation; the performance is investigated using the mean opinion score(MOS) test and delay in the presence of AWGN channel.**

1 INTRODUCTION

Speech signal is the most important one helping people to communicate; although people can transfer ideas through face emotions, body movements and written material, but the speech signal is the fastest and easiest way to transfer ideas.

One of the most important and economical speech processing applications is the speech coding. The speech is segmented into short segments (20 ms for example) and the encoder extracts some features of the speech signal, and sends the features instead of the signal itself. At the receiver the signal is reconstructed from these features.

The world of technology has given mankind a powerful way for interaction using Telecommunication. When invented by Alexander Graham Bell, it was a wired transmission of electrical signals representing information. Since then, telecommunication technology has achieved tremendous improvement from text, voice transmission to a modern age high speed real time multimedia content. The challenges for today's technology is to develop standards that can help operators to keep the cost per bit as low as possible, maintaining backward compatibility so as to gain maximum benefit from the investments. Newer modulation schemes and improved advanced antenna technologies are helping to achieve the newer heights of success. The technology so far has developed through 1st, 2nd and 3rd generation phases and currently 4G (4th Generation) is the best experience till date for users [1].

Two emerging technologies, the IEEE 802.16 WiMAX (Worldwide Interoperability for Microwave Access) and the 3GPP LTE (Third Generation Partnership Project Long Term Evolution) aim to provide mobile voice, video and data services by promoting low cost deployment. Both technologies are intended to offer ubiquitous broadband at multiple megabits per second.

The rest of the paper is organized as follows: a brief about WiMAX and LTE is introduced in section 2. Section 3 describes speech coding and different types of speech coders. Section 4 includes the simulation models. Section 5 provides the measurement tools.Section6providesthe results of the experimental work and simulation results. Finally, section 7 concludes the paper.

2 MOBILE SYSTEMS

The International Telecommunications Union-Radio communications sector (ITU-R) specified a set of requirements for 4G standards, named the IMT-Advanced (International Mobile Telecommunications Advanced) specification, setting peak speed requirements at 100 megabits per second (Mbit/s) for high mobility communication (such as from trains and cars) and 1 gigabit per second (Gbit/s) for low mobility communication. To meet IMT-Advanced requirements, IEEE 802.16e (Mobile WiMAX) an

IEEE standard and LTE from 3GPP groups are considered and both satisfy the IMT Advanced requirements [1].

WiMAXis a technology which provides wireless transmission of data in variety of ways, ranging from point to point links to the full mobile cellular access. The WiMAX physical layer is based on orthogonal frequency division multiplexing a scheme that offers good resistance to multipath, and allows WiMAX to operate in non-line of sight (NLOS) conditions. MOBILE WiMAX uses Orthogonal Frequency Division Multiplexing(OFDM) as a multiple access technique, whereby different users can be allocated different subsets of the OFDM tones. OFDMA facilitates the exploitation of frequency diversity and multiuser diversity to significantly improve the system capacity [2]. It adopts different modulation and coding schemes (MCS), hence follows adaptive modulation and coding (ACM) scheme as the received signal strength of a user varies in a cell. Initially the operation band of frequencies was 10-66GHz which was line of sight (LOS) and in the later amendments 2-11GHz band is used which is (NLOS) [3].

LTE was developed in the Third-Generation Partnership Project as the natural progression of High-Speed Packet Access (HSPA).LTE is a modulation technique that is designed to deliver 100Mbps per channel and give individual users performance comparable to today's wired broadband, and it uses OFDM in downlink and Single Carrier-Frequency Division Multiple Access (SC-FDMA) in uplink [4]. It promises high peak data rates for uplink and downlink transmission, spectral efficiency, low delay and latency, low bit error rates [5].

3 SPEECH CODING

Speech coding is the process of obtaining a compact representation of voice signals for efficient transmission over band-limited wired and wireless channels and/or storage capacity. Today, speech coders have become essential components in telecommunications and in the multimedia infrastructure. Commercial systems that rely on efficient speech coding include cellular communication, voice over internet protocol (VOIP) and videoconferencing.

Most telecommunications coders are *lossy*, meaning that the synthesized speech is perceptually similar to the original but may be physically dissimilar.

Speech coders differ primarily in bit rate (measured in bits per sample or bits per second), complexity (measured in operations per second), delay (measured in milliseconds between recording and playback), and perceptual quality of the synthesized speech [6].

There are different types of speech coders,

A. Waveform Coders

Attempt to code the exact shape of the speech signal waveform, without considering the nature of human speech production and speech perception. These coders are high-bit-rate coders (typically above 16 kbps).

B. Vocoders

Preserve only the spectral properties of speech in the encoded signal.
Vocoders produce intelligible speech at much lower bit rates, but the level of speech quality in terms of its naturalness and uniformity for different speakers is also much lower. The applications of vocoders so far have been limited to low-bit-rate digital communication channels. The linear predictive coding (LPC) vocoders which are based on the speech production model operate at bit rates as low as 2kbits/s. However, the synthetic quality of the vocoded speech is not broadly appropriate for commercial telephone applications.

C. Hybrid Coders

The main limitation of the LPC vocoding is the assumption that speech signals are either voiced or unvoiced, hence the source of excitation of the synthesis all-pole filter is either a train of pulses (for voiced speech), or random noise (for unvoiced speech).In fact there are more than two modes for which vocal tract is excited and often these modes are mixed. Hybrid coders combine features from both waveform coders and vocoders. Several hybrid coders employ an analysis-by-synthesis process in order to derive code parameters.

The speech coders that will be simulated and implemented in our work are:
**G.711*
G.711 is a public domain codec widely used in VoIP applications. It was introduced in 1972 by the ITU. It employs a logarithmic compression that compresses each 16-bit sample to 8-bits. As a result, its bit-rate is 64 kbps, which is considered the highest bit-rate among the codecs. G.711 offer very good speech quality [7].

***** *G.729*
G.729 is a licensed codec designed to deliver good call quality without consuming high bandwidth [7]. It is based on the Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP) algorithm with bit-rate of 8 kbps [8]. The coder operates on speech frames of 10 ms corresponding to 80 samples. In addition, there is a look-ahead of 5 ms, resulting in a total algorithmic delay of 15 ms [9].

* *G.723.1*
G.723.1 is also a licensed codec. It is designed for calls over modem links with data-rates of 28.8 and 33 kbps. Therefore, it has two versions with distinct bit-rates: 5.3 and 6.4 kbps [7]. In this paper, we consider the 6.4 kbps, which is based on the Multi-Pulse Maximum Likelihood Quantization (MP-MLQ). This coder encodes speech or other audio signals in 30 msec frames. In addition, there is a look ahead of 7.5 msec, resulting in a total algorithmic delay of 37.5 msec [10].

**iLBC*
The internet Low Bit rate Codec (iLBC) is suitable for robust voice communication over IP. The codec is designed for narrow band speech and results in a payload bit rate of 13.33 kbps with an encoding frame length of 30 ms and 15.20 kbps with a frame length of 20 ms. The iLBC codec enables graceful speech quality degradation in the case of lost frames, which occurs in connection with lost or delayed IP packets [11], we consider the 13.33 kbps which is based on block-independent linear predictive coding.

4 SIMULATION OF THE IMPLEMENTED SYSTEM

Figure 1 shows a general block diagram for the transmission of the speech codecs over mobile wireless system, all simulations were carried out using Matlab R2009a, Matlab built-in function, “awgn” used to generate AWGN channel.
We have implemented only the mandatory features of the physical layer of WiMAX and LTE, and assumed a single input single output (SISO) scenario.

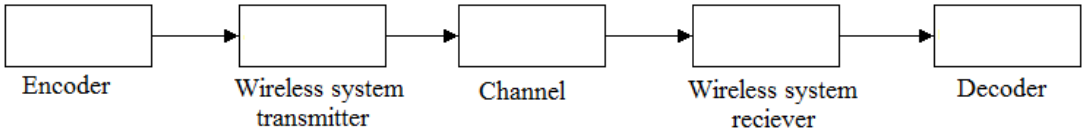
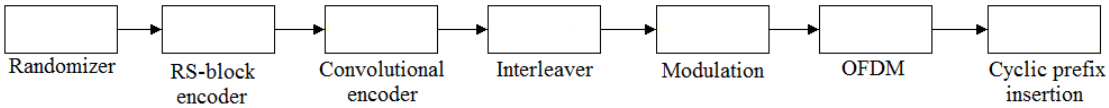
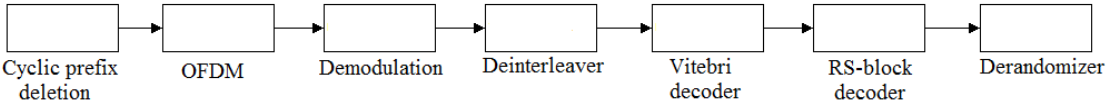


Figure 1: General block diagram for the speech transmission over wireless system.

Figure 2 shows the block diagram of the mobile WiMAX transceiver.



(a) WiMAX transmitter

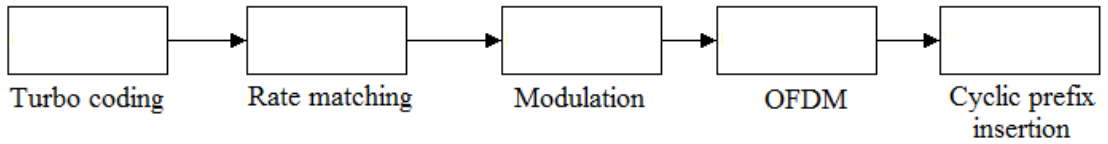


(b) WiMAX receiver

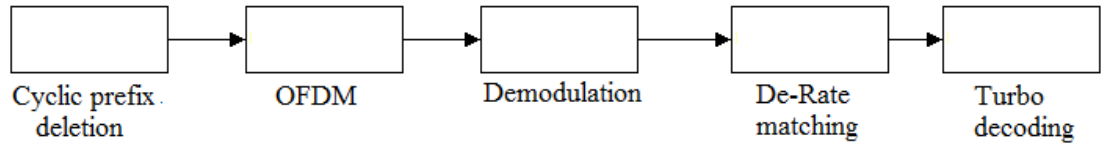
Figure 2: Block diagram of WiMAX physical layer

The randomization process has been carried out to scramble the data in order to convert long sequences of 0's or 1's in a random sequence to improve the coding performance. After that, we have performed Reed-Solomon (RS) encoding with the parameters (N = 255, K = 239, T = 8). 2/3 rated convolutional encoding is also implemented separately on the RS encoder. The encoding section was completed by interleaving the encoded data. The QPSK digital modulation technique is then used to modulate the encoded data. The modulated data in the frequency domain is then converted into time domain data by performing IFFT on it. For reducing inter-symbol interference (ISI) cyclic prefix (CP) has been added with the time domain data, CP is generated by duplicating the last G samples of IFFT output symbol and adding them to the beginning of that symbol. Finally the modulated parallel data were converted into serial data stream and transmitted through the communication channel. At the receiving section, the reverse procedures of the transmission section have been performed.

Figure 3 shows a general block diagram for LTE physical layer



(a) LTE transmitter side



(b) LTE receiver side

Figure 3: Block diagram of LTE physical layer

The input bits are compared to the maximum code word size which is 6144 bits and if larger, the segmentation is performed. Segmented code blocks are coded using a turbo coder with the rate of 1/3. The rate matching block consists of three stages, a sub block interleaver (each code word is interleaved individually), bit collection, and then bit selection and puncturing. The modulation mapper modulates the bits with QPSK scheme and the modulated data in the frequency domain is then converted into time domain data by performing IFFT on it. For reducing inter-symbol interference (ISI) cyclic prefix has been added with the time domain data, finally the modulated parallel data were converted into serial data stream and transmitted through the communication channel. At the receiving section, the reverse procedures of the transmission section have been performed.

5 MEASURING THE PERFORMANCE OF SPEECH QUALITY

A. MOS Performance

The 'gold standard' for measuring voice quality, i.e. 'listening quality' is the mean opinion score (MOS) as specified by ITU-T recommendation P800 [R.2].

The human listeners are required to classify the perceived voice quality into categories (excellent, good, fair, poor and bad) and a standardized averaging process is applied to produce a final 'MOS score' from a large number of independent assessments. A definition of the MOS categories as presented to volunteers and their scores is given in Table I.

TABLE I
DEFINITION OF MOS SCORE

Rating	Voice quality	Level of distortion
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible & slightly annoying
2	Poor	Annoying but not objectionable
1	Bad	Very annoying and objectionable

The averaging process produces scores on a continuous scale between one and five, and an average score of 4.0 or higher is referred to as 'toll quality', this being, in principle, the quality of speech received over a normal domestic wired PSTN telephone link. Such a link would be 'narrow-band' (300-3.4kHz) with speech sampled at 8 kHz and normally transmitted by A-law PCM (G711) at 64 kb/s.

B. Delay Performance

Delay is also one of the more important aspects in measuring quality of speech.
The delay of the voice over codecs and the delay of the voice over codecs over wireless systems will be investigated.
The processor used in our simulation is Intel Core™ i5-560M Processor 2.66 GHz.

6 SIMULATION RESULTS

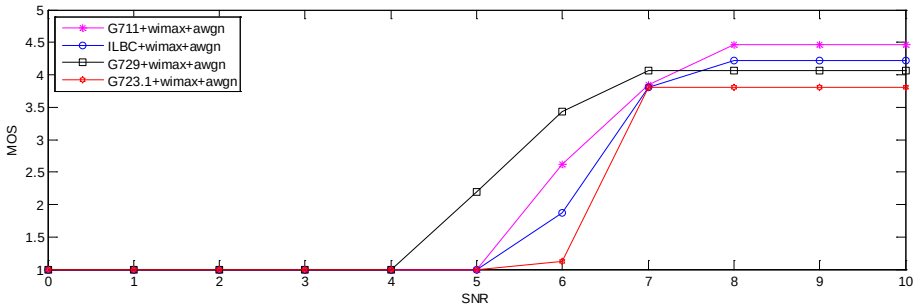
Table II shows the results of the MOS test without noise of the four codecs

TABLE II
MOS OF VOICE CODERS

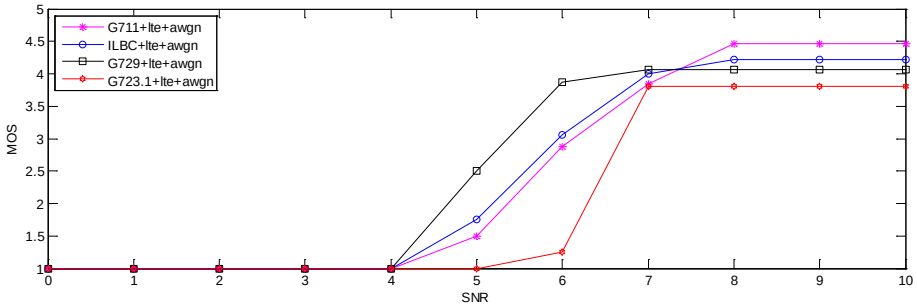
Codec	MOS
G.711	4.4
iLBC	4.2
G.729	3.95
G.723.1	3.8

From the above table it's clear that G.711 codec has the best MOS in the absence of noise since it has the highest bit rate followed by iLBC and G.729 coders.

Figure 4 shows the MOS of the four codecs over the WiMAX and LTE at different signal to noise (SNR) ratios.
In case of WiMAXG.729 has the best MOS in the presence of AWGN followed by G.711 and iLBC.
In case of LTE G.729 has the best MOS in the presence of AWGN followed by iLBC and G.711.



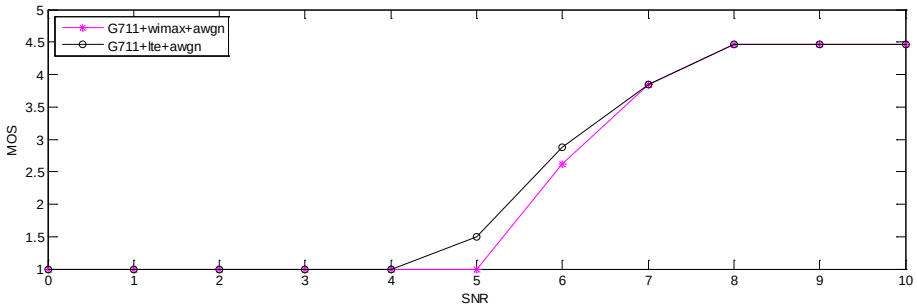
a) Results of MOS test of the four codecs over WiMAX.



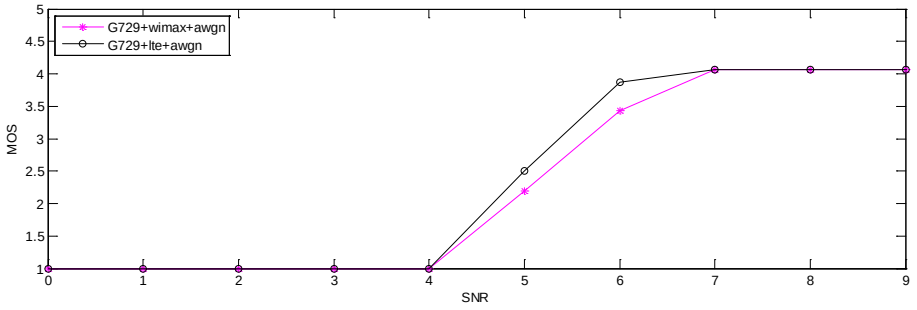
b) Results of MOS test of the four codecs over LTE.

Figure 4: MOS of the four codecs over WiMAX and LTE
a) WiMAX b) LTE

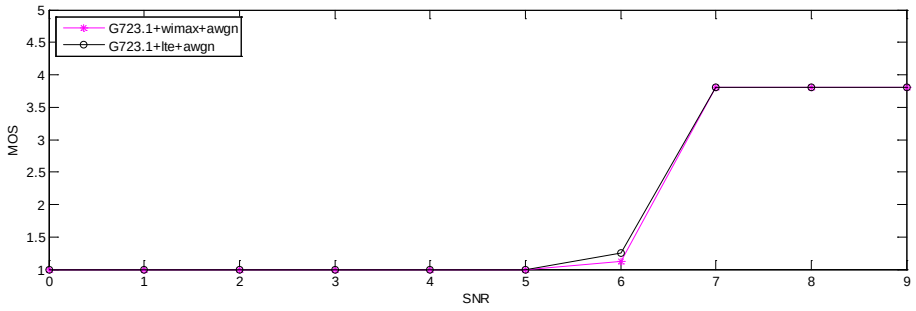
Figure 5 shows the MOS of the different codecs over the WiMAX and LTE wireless systems. All codecs over LTE has a better MOS in the presence of AWGN than over WiMAX, and this is a logical result since LTE uses a turbo coding while WiMAX uses convolutional coding.



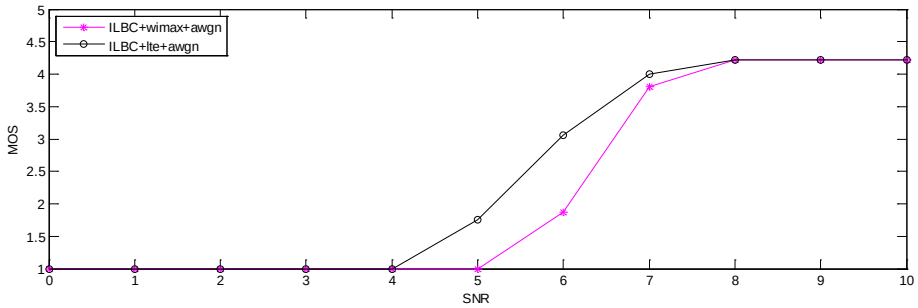
a) MOS of G.711 codec over WiMAX and LTE



b) MOS of G.729 coder over WiMAX and LTE



c) MOS of G.723.1 coder over WiMAX and LTE



d) MOS of ILBC over WiMAX and LTE

Figure 5: MOS of the coders over WiMAX and LTE
a) G.711 b) G.729 c) G.723.1 d) iLBC

In table III the delay of the voice over codecs and the delay of the codecs parameters over wireless systems are shown.

TABLE III
DELAY

Codec	Codec delay (s)	Total delay over WiMAX (s)	Total delay over LTE (s)
G.711	0.01	2.25	0.08
ILBC	0.0147	2.11	0.06
G.723.1	0.07	2.01	0.02
G.729	0.0156	1.05	0.01

From the above table it's clear that G.729 codec has the lowest delay over both wireless systems followed by G.723.1 and iLBC codecs since it processes small frames.

7 CONCLUSIONS

In this paper, we evaluated the performance of four voice codecs over WiMAX and LTE under AWGN channel, and we will evaluate the performance under Rayleigh fading channel later. The MOS and total end-to-end delay were used as performance parameters. G.711 codec has the highest MOS in the absence of noise due to its high bit rate but G. 729 codec offers the best performance over both WiMAX and LTE.
The LTE wireless system has a lower delay (since it has a lower complexity) and a better MOS (due to the presence of a turbo coder in its physical layer) than WiMAX wireless system. However, all four codecs G. 711, G. 729, G. 723.1, and ILBC show acceptable performance quality for voice over WiMAX and LTE.

REFERENCES:

[1] S.S. Pisal, "Physical Layer Comparative Study of WIMAX and LTE", Master thesis, San Diego State University, Spring 2012.
[2] C.M. Babu and H. Rana, "Estimating the Channel with Different Fading Characteristics and to Develop an Algorithm to Minimize the Effect of Fading for Broadband Services in Multiple Path Communication in Wi-Max," *International Journal of Computer Networks and Wireless Communications (IJCNWC),vol.2, no.2,ISSN: 2250-3501, April 2012.*

[3] S. Kulkarni, "Performance Evaluation of VoIP in Mobile WiMAX; Simulation and Emulation studies," *International Journal on Computer Science and Engineering (IJCSE), Vol. 3, no. 3, ISSN: 0975-3397, Mar 2011*.
[4]F. Rezaei, "A Comprehensive Analysis of LTE Physical Layer", Master thesis, University of Nebraska, December 2010.
[5] G. M. Kebede and O. O. Paul, "Performance Evaluation of LTE Downlink with MIMO Techniques", Master thesis, Blekinge Institute of Technology, Sweden, November 2010.
[6] IEEE802.16e: IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, 2005.
[7] V. Toncar, "VoIP basics overview of audio codecs", April 2012, Available:
http://toncar.cz/Tutorials/VoIP/VoIP_Basics_Overview_of_ Audio_Codecs.html
[8] VoIP codecs. April 2012. Available:
http://www.voip-sip.org/voip-codecs
[9] ITU-T Recommendation G.729: "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)", 2007.
[10] ITU-T Recommendation G.723.1: "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s", 1996.
[11] A.A. Ali , S. Vassilaras and K. Ntagkounakis," A Comparative Study of Bandwidth Requirements of VoIP Codecs Over WiMAX Access Networks", Third International Conference on Next Generation Mobile Applications, Services and Technologies, 2009.NGMAST '09. pp. 197-203, Cardiff, Wales.

قياس كفاءة بعض مشفرات الصوت المختلفة على Wimax, LTE

نعمت سيد عبد القادر، نرمين رسمى، هبة الله مصطفى مراد

كلية الهندسة – جامعة القاهرة

تم فى هذا البحث مقارنة جودة الصوت الناتج من مشفرات صوت مختلفة من خلال جيلين من أجيال الأتصالات المتحركة Wimax, LTE.

ولقد تمت محاكاة **ITU-T G.711, ITU-T G.729, ITU-T G.723.1 and iLBC** مشفرات الصوت الأتية:

ولقد تم قياس كفاءة المشفرات باستخدام اختبار الرأى المتوسط وكذلك بقياس التأخير فى انتاج الصوت فى وجود قناة **AWGN.** .

ولقد أثبتت النتائج تفوق iLBC. على باقى المشفرات وعموما كانت جودة الصوت الناتج من جميع المشفرات جيدة.

- The main objective of our research is to prove that using GA with CRF may improve the results of using CRF only.
- We have already achieved this objective

- Although using GA with CRF improves the results it takes several hours to run

7 CONCLUSION AND FUTURE WORK

Nowadays researchers turned from the age of building systems using single classifiers to the age of mixing the learning techniques to build more effective systems.
The integration of machine learning techniques becomes like the art. We proved that we can easily use CRF with GA together in order to build a simple Semi-supervised machine learning Arabic named entity recognition task using a simple feature setup. To our knowledge CRF and GA aren't used together before in this field.
The next step comes after this research is to expand the system to cover more entity types, also we intend to use search algorithms to select the best feature set suitable for these entities in order to increase the F-measure. On other word GA can be also used to select the feature set in our work.And this may enhancethe obtained results.

REFERENCES

[1] Y.Benajiba, and P.Rosso,"Arabic Named Entity Recognition using Conditional Random Fields", *Conference on Language Resources and Evaluation(LREC)*,May, 2008.
[2] D.Nadeau and S.Sakine," A survey of named entity recognition and classification", *National Research Council* Canada,New York,USA,2006.
[3] S. AbdelRahman, M. Elarnaoty, M. Magdy and A. Fahmy,"Integrated Machine Learning Techniques for Arabic Named Entity ", *International Journal of Computer Science Issues(IJCSI)*, Vol. 7, Issue 4, No 3, July 2010.
[4] A.Farghaly and K.Shalaan,"Arabic Natural Language Processing:Challenges and Solutions ,*ACM Journal* , Vol. 8,issue 4,No 14,New york,December,2009.
[5] M.Wallach,"Conditional Random Fields an introduction", *Technical Report (CIS), University of Pennsylvania*,February, 2004.
[6] A.Abdelhameed andK.Darwesh,"Simplified feature set for Arabic named Entity Recognition",*Proceedings of the 2010 Named Entities Workshop*,pp110–115, Uppsala, Sweden, 16 July 2010.
[7] D.Whitly,"A Genetic Algorithm tutorial".*Statistics and Computing*, volume 4,pp 65-85,Colordo State University", 1994.
[8] M.yuh day,C.Hung,C.Shyong and S.Humg,Integrating Genetic Algorithms with Conditional Random Fields to EnhanceQuestion Informer Prediction,*IEEE Intelligent Systems*,volume 13,pp 44-49,1998.

دمج "CRF " مع "GA" لبناء نظام التعرف على الكيانات العربية المسماة نصف شبه موجه

نهى أحمد سعد الدين*[1], على فرغلي**[2], على على فهمي*[3]
*كلية الحاسبات والمعلومات ,جامعة القاهرة
[1]nohaahmed_fci@yahoo.com
[3]Aly.fahmy@cu.edu.eg
**Text Group, Oracle USA, Redwood Shores, CA.
[2]alifarghaly@yahoo.com

ملخص:
اصبحت مهمة البحث عن الكيانات المسماه(NER) في الوقت الحاضر مهمه فرعيه ضروريه جدا للعديد من مهام معالجة اللغات الطبيعيه لأنها تحسن الى حد كبير ادائها,تقدم هذه الورقه البحثيه حل لتنفيذ هذه المهمه في اللغه العربيه(ANER)ويقوم الحل على أساس التعليم الشبه موجه للحاسب وهو عباره عن تكامل واندماج بين استخدام ال "CRF"باعتباره مصنف يمكن استخدامه في تصنيف الكيانات و الخوارزميات الجينيه (GAs) والتي تستخدم في البحث عن افضل الحلول.
تستخدم مساهمات هذا العمل العلميه في فكرة استخدام وسيلة بسيطة لبناءنظام هجين ANER بدءا من اختيارمجموعة ميزة بسيطة لتطويرخوارزمية شبه موجهه بسيطة.لم يقترح هذاالحل المقترح ل ANER من قبل.
الحل يثبت أن استخدام GA جنبا إلى جنب معCRF يتفوقCRF على العمل الوحيد

Graph Reduction in Abstractive Text Summarization

Marwa Mahmoud[1], Ibrahim Fathy[2], Mostafa Aref[3]

Department of Computer Science,
Faculty of Computer and Information Sciences,
Ain-Shams University, Cairo, Egypt.
[1]marwa_mahmoud100@hotmail.com
[2]ibrahim.moawad@heic.eg
[3]mostafa.m.aref@gmail.com

***Abstract*— One of the important Natural Language Processing applications is Text Summarization, which helps users to manage the vast amount of information available, by condensing documents' content and extracting the most relevant facts or topics included. Text Summarization can be classified according to the type of summary: extractive, and abstractive. Extractive summary is the procedure of identifying important sections of the text and producing them verbatim while abstractive summary aims to produce important material in a new generalized form. In this paper, several approaches for graph reduction are presented. These approaches have different applications in different fields. The paper describe in progress research to apply graph reduction techniques in abstractive text summarization.**

1. INTRODUCTION

Informally, a graph is set of nodes, pairs of which might be connected by edges. In a wide array of disciplines, data can be intuitively cast into this format. For example, computer networks consist of routers/computers (nodes) and the links (edges) between them. Social networks consist of individuals and their interconnections (which could be business relationships or kinship or trust, etc). Protein interaction networks link proteins which must work together to perform some particular biological function. Ecological food webs link species with predator-prey relationships. In these and many other fields, graphs are seemingly ubiquitous [1]. The reduction [2], or simplification, of graph-based models is critical to the analysis, simulation and control design for systems arising in many diverse areas such as network routing, image processing, statistical learning, and in distributed control of networked dynamical systems, to name a few. The reduction is carried out through node and edge aggregation, where the simpler graph is representative of the original large graph.

The primary aim of our research is to systematically select and review published work and provide an overview of graph reduction theory, their methods, applications and focuses. This paper proceeds as follows. We provide a description of relevant reviewed papers and classify them into appropriate categories according to topics. Typically, we only provide the main ideas and approaches; the interested reader can read the relevant references for details. In all of these, we attempt to collate information from several fields of research. Our conclusions are presented in the last section.

2. SELECTED STUDIES

We identified five main categories for graph reduction applications: workflow management system, computer vision, networks, semantic graphs and other application.

A. Workflow Management System

Workflow technology has been a new hotspot in the area of computer application since 1990. Nowadays workflow management systems are widely used in improving the effectivity and efficiency of business processes. The production workflows, a subclass of workflows, support well-defined procedures for repetitive processes and provide a means for automated coordination of activities that may span over several heterogeneous and mission-critical information systems of an organization [3]. Production workflow applications are built upon business processes that are generally quite complex and involve a large number of activities and associated coordination constraints. Using a generic process modeling language for workflows, we show how a structural specification may contain deadlock and lack of synchronization conflicts that could compromise the correct execution of workflows. It is essential that a process model is properly defined, analyzed, verified, and refined before being deployed in a workflows management system. Sadiq and Orlowska in [3] presented an effective graph reduction algorithm that can detect the existence of structural conflicts in workflow graphs. The basic idea behind verification approach is to remove some nodes and/or transitions from the workflow graph G. and keeping the reduction process as long it can reduce the size of process graph. The algorithm

reduces a workflow graph without structural conflicts to an empty graph. it means that the original process graph G before reduction does not contain any structural conflicts. Otherwise, it contains deadlock or lack of synchronization structural conflicts that would be clearly visible form the reduced graph. The algorithm in [4] is based on a set of graph reduction rules to identify the deadlock and lack of synchronization conflicts that could compromise the correct execution of a workflow. A complete set of graph reduction rules is presented to reduce the workflow graph. The graph reduction algorithm can remove all nodes from workflow graphs that are definitely correct. In [5] Lu and Liu proposed a combined graph-reduction and graph search algorithm which is called "CWRS" to verify workflow graphs. This algorithm reduces a cyclic workflow graph to an acyclic workflow graph, and then verifies this acyclic workflow graph. The computational complexity of our algorithm behaves a good performance. So it should be valuable in practical graph-based workflow verification.

B. Computer Vision

Graph cuts had a growing impact in shape optimization. In particular, they are commonly used in applications of shape optimization such as image processing, computer vision and computer graphics. Solving problems with a large number of variables remains computationally expensive and requires a high memory usage since underlying graphs sometimes involve billion of nodes and even more edges. Examples for computer vision problems include segmentation, image restoration, dense stereo estimation and shape matching. An efficient algorithm for image segmentation using GraphCuts which can be used to efficiently solve labeling problems on high resolution images or resource-limited systems is presented in [6]. A Slim Graph is constructed by merging nodes that are connected by simple edges. A proof is given that the value of the maximum flow on the Slim Graph is equal to the maximum flow of the original graph. The nodes connected by a simple edge will have the same label in the final segmentation and can be merged into a single node. Thus the original graph is simplified to a Slim Graph without changing the energy-minimization problem and the value of the global minimum. It was shown that the proposed method required much less memory allowing segmentation of images of reasonable sizes even on mobile devices. A further reduction of computation time can be achieved by using parallel hardware architecture. In [7] Malgouyres and Lerme proposed a method to improve graph-cuts in this regards. A formal statement is given which expresses that a simple and local test performed on every node before its construction permits to avoid the construction of useless nodes for the graphs typically encountered in image processing and vision. A useless node is such that the value of the maximum Flow in the graph does not change when removing the node from the graph. Such a test therefore permits to limit the construction of the graph to a band of useful nodes surrounding the final cut. Energy-minimizing active contour models (snakes) have been proposed for solving many computer vision problems such as object segmentation, surface reconstruction, and object tracking. Energy-minimizing active contour models (snakes) have been proposed for solving many computer vision problems such as object segmentation, surface reconstruction, and object tracking [8]. Dynamic programming which allows natural enforcement of constraints is an effective method for computing the global minima of energy functions. However, this method is only limited to snake problems with one dimensional (1D) topology (i.e., a contour) and cannot handle problems with two-dimensional (2D) topology. Yan et al. [8] presented an algorithm to minimize the energy function associated with 2D snakes. 2D snakes represent 2D surfaces with connected deformable graphs controlled by vertices and edges. A set of reduction operations are defined and used to simplify the graph of the 2D snake into one single vertex while retaining the minimal energy of the snake.

C. Networks

Dynamic, QoS-based routing received considerable attention in recent years, especially considering the difficulty in predicting Internet traffic patterns and the consequent impossibility to properly plan and dimension the network. The core of any QoS-based routing algorithm is a network status- dependent cost function that is used to find the optimal (or at least a suitable) route across the network by solving an optimization problem. Routing can be formalized as the problem of finding a suitable set of edges connecting two nodes in a directed graph. Casetti et al. [9] introduced a new approach to QoS routing, the approach can be summarized as NGR (Network Graph Reduction) i.e., a modification of the graph describing the network before the routing path is computed, in order to exclude from the path selection over-congested portions of the network. This solution leads to a class of two-step routing algorithms, where both steps are simple, hence allowing efficient implementation. Chekuri and Korula [10] gave two applications of a graph reduction step to connectivity and network design problems. This step preserves the global and local connectivity of the graph. The first, is a polylogarithmic approximation for the problem of packing element-disjoint Steiner forests in general graphs, and an O(1)-approximation in planar graphs. The second is a very short and intuitive proof of a spider-decomposition theorem of

Chuzhoy and Khanna [11] in the context of the single-sink k-vertex-connectivity problem. The results highlight the effectiveness of the element-connectivity reduction step.

D. Semantic Graphs

The increasing availability of online information has necessitated intensive research in the area of automatic text summarization within the Natural Language Processing (NLP) community. Automatic text summarization can be done using a semantic graph reducing technique. The reader thus obtains an overview of the content, without having to read through the text. In building a compact semantic graph, an important step is grouping similar concepts under the same label and connecting them to external repositories. Few papers address the problem of reduction and enhancement of semantic graphs. A new method [12] is proposed for reducing large directed graphs to simpler graphs with fewer nodes. The reduction is carried out through node and edge aggregation based on the maximum entropy principle. As a special case, this method applies to the Markov chain model-reduction problem, providing a soft-clustering approach that enables better aggregation of state-transition matrices than existing methods. Graph reduction is utilized for efficient and effective indexing and retrieval. In [13] the authors focus on the role of semantic graph in web page content visualization and the role of graph in displaying semantic relations. The semantic graph is generated in the form of subject, object and verb where subject and object are represented by nodes and verb defines the relationship between them. It also includes the overall process of a system in creating a smaller semantic graph from the given web page html document. More compact semantic graphs can be generated [14] by identifying the triplet elements that share the same meaning and can be therefore merged together under the same label. Additionally, linking semantic graph nodes to external resources, such as WordNet thus helping in better understanding the graph content. Tamil Document Summarization using sub graph [15] presents a method for extracting sentences from an individual document to serve as a document summary or a pre-cursor to creating a generic document abstract. Semantic features of the text are identified using Logical Form (LF) Parser. The semantic graph constructed using the LF parser is then used to select important parts of the document for summary generation. The selection of important sentences from the graph is based on three types of attributes, Linguistic attributes, Semantic Graph attributes and Document Discourse Structure attributes considered as features for learning. This rich set of features serves as input to Support Vector Machine (SVM) classifier which classifies the sentences as important or unimportant for inclusion into the summary. In [16] a novel approach is presented to create an abstractive summary for a single document using a rich semantic graph reducing technique. A model of heuristic rules is applied to reduce the graph by replacing, deleting, or consolidating the graph nodes using the WordNet relations.

E. Other Applications

Reduction is matching invariant, so it can be used as a speed-up in matching algorithms. Bartha and Kresz [17] presented a linear-time algorithm to shrink a graph G recursively along its 2-star subgraphs called redexes. The starting point was a greedy algorithm, which works in linear time only if G does not contain recursively incurring (implied) redexes. Implied redexes have been detected and reduced during a single bottom-up sweep of the depth-first tree of G, and the resulting graph was transferred to the greedy algorithm to construct the desired graph r(G).

Pointer manipulation is notoriously dangerous in languages like C where there is nothing to prevent: the creation and dereferencing of dangling pointers; the dereferencing of nil pointers or structural changes that break the assumptions of a program, such as turning a list into a cycle. Graph reduction specifications (GRSs) are a powerful new method for specifying classes of pointer data structures (shapes). They cover important shapes, like various forms of balanced trees that cannot be handled by existing methods. Bakewell et al. [18] showed how to improve the safety of pointer programs by providing (1) means for programmers to specify pointer data structure shapes by graph reduction specifications (GRSs), which are the dual of graph grammars in that graphs in a language are reduced to an accepting graph rather than generated from a start graph, and (2) algorithms to check statically whether programs preserve the specified shapes.

Natural Language Understanding (NLU) technology is a fundamental component of dialog-based automatic speech understanding systems. Such systems are typically implemented on telephony platforms and are used to automate the communication process between humans and machines through natural speech. An important component of such systems is the semantic parser whose purpose is to recognize structures in the sentence, with the goal to facilitate meaning extraction. In [19] Huertu and Lubensb introduced a method to represent the semantic parser domain into a single directed graph showing the parser's labels and their immediate inter-relationships as they exist in the annotated

development corpora. The authors describe how the graph representation method can be utilized in the reduction of the complexity of the parser by identification and removal of nodes, edges and structures of the domain graph without major impact in the parser's accuracy.

Modern embedded systems typically consist of both hardware and software components. Hardware/software (HW/SW) partitioning Hardware/software (HW/SW) partitioning is one of the key processes in an embedded system. It is used to determine which system components are assigned to hardware and which are processed by software. The design's performance is directly determined by the results of HW/SW partitioning. In [20] Hui et al. presented the idea of graph reduction techniques for HW/SW partitioning. The task graph reduction technology, which is proposed for HW/SW partitioning, consists of the sub-graph searching algorithm and the Sub-graph evaluation and selecting algorithm. The purpose of this pre-process for the task graph is to find all the sub-graphs and to reduce each sub-graph to a single task node, which results in a new reduced task graph. The partitioning algorithm for the reduced task graph achieved improvement on partitioning speed and accuracy and also is able to make full use of the hardware area.

Goldman and Ngoko investigated the question of Service Response Time (SRT) prediction of a Web Service Composition using the graph reduction technique [21]. The authors propose a fast algorithm for graph reduction that uses less memory and perform fewer operations. The graph reduction approach proceeds in two phases. The first phase deals with the computation of a reduction order. The idea in such an order is to define an ordered set of decomposable subgraphs for the reduction. The second stage is the reduction with the defined order.

The bug localization techniques based on graph mining are successfully applied in a wide range of practical problems arising in software industry. One of the techniques for automated bug localization is usage of call graph. Since size of the call graph generated is quite large, in [22] a novel algorithm for call graph reduction has been proposed in order to use the respective call graphs for bug localization, the developed technique stores the parent information in the matrix and reduced at each level drastically. Information about each node is retained by using the call frequency by annotating each edge with a numerical weight. Similarly the algorithm used to reduced call graph has various advantages over traditional techniques.

Consider an agent who seeks to traverse the shortest path in a graph having random edge weights. If the agent has no information about the realizations of the edge weights, it should simply take the path of least average length (a deterministic optimization). Generalizations of this framework whereby the agent has access to a limited amount of side information about the edge weights ahead of choosing a path. Rinehart and Dahleh [23] presented A new graph reduction for analyzing the value of side information for shortest path optimization was presented this graph reduction that captures limited but significant information about the geometry of the graph's path polytope. The resulting reduction serves to concentrate side information to its ability to help the agent discern the shortest path in the graph, not simply determine its length.

3. CONCLUSION

This paper presents several graph reduction techniques. These techniques and their applications in different fields are discussed. These techniques include Workflow Management System, Computer Vision, Networks and semantic graphs. The objective of this research is to apply graph reduction techniques in abstractive text summarization.

REFERENCES

[1] Deepayan Chakrabarti, Christos Faloutsos: *Graph mining: Laws, generators, and algorithms*. ACM Comput. Surv. 38(1): (2006)

[2] Y. Xu, S.M. Salapaka, and C.L. Beck, "On reduction of graphs and Markov chain models", in Proc. CDC-ECE, pp.2317-2322, 2011.

[3] Wasim Sadiq and Maria E. Orlowska," Analyzing Process Models Using Graph Reduction Techniques" Information Systems, 25(2): 117-134, 2000.

[4] H. Lin, Z. Zhap, H. Li, and Z. Chen, "A novel graph reduction algorithm to identify structural conflicts," in Proc. 35th Hawaii Int. Conf. Syst. Sci., vol. 9, p. 289, 2002.

[5] Kai Lu; Liu, Qiang, "An Algorithm Combining Graph-Reduction and Graph-Search for Workflow Graphs Verification". En: 11th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2007. Melbourne, Australia: Springer, p. 772-776, 2007.

[6] B. Scheuermann and B. Rosenhahn, "SlimCuts: GraphCuts for High Resolution Images Using Graph Reduction," Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR), July 2011.

[7] F. Malgouyres and N. Lermé, "A non-heuristic reduction method for graph cut optimization," Technical Report hal-00692464, CCSD, April *2012*.

[8] Yan, J., K. Zhang, Z. Zhang, and S.C. Chen, "A Graph Reduction Method for 2D Snake Problems," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, p. 6, 2007.

[9] Casetti, C.; Lo Cigno, R.; Mellia, M.; Munafo, M., "A new class of QoS routing strategies based on network graph reduction," INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, vol.2, no., pp.715,722 vol.2, 2002.

[10] C. Chekuri and N. Korula," A graph reduction step preserving element-connectivity and applications," In ICALP (1), pages 254–265, 2009.

[11] Chuzhoy, J., Khanna, S.,"Algorithms for Single-Source Vertex-Connectivity," In: Proc. of IEEE FOCS, pp. 105–114 (October 2008)

[12] Y. Xu, S.M. Salapaka, and C.L. Beck, "On reduction of graphs and Markov chain models", in Proc. CDC-ECE, pp.2317-2322, 2011.

[13] Sushil Shrestha, "Role Of Semantic Graph In Web Page Content Visualization". Athmandu University Journal Of Science, Engineering And Technology Vol. 8, No. I, 125- 133, February, 2012.

[14] Delia Rusu, Blaz Fortuna, Dunja Mladenić " Improved Semantic Graphs with Word Sense Disambiguation". Jozef Stefan Institute, Semantic Web, kcap09.stanford.edu, 2009.

[15] Banu, M.; Karthika, C.; Sudarmani, P.; Geetha, T.V., "Tamil Document Summarization Using Semantic Graph Method," Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on , vol.2, no., pp.128,134, 13-15 Dec. 2007

[16] Moawad, I.F.; Aref, M., "Semantic graph reduction approach for abstractive Text Summarization," Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on , vol., no., pp.132,138, 27-29 Nov. 2012

[17] Bartha, M.; Kresz, M., "A Depth-first Algorithm to Reduce Graphs in Linear Time," Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2009 11th International Symposium on , vol., no., pp.273,281, 26-29 Sept. 2009.

[18] A. Bakewell, D. Plump, C. Runciman." Specifying Pointer Structures by Graph Reduction,"
In AGTIVE 2003, pp. 30–44. Springer-Verlag, 2004.

[19] Huerta, J.M.; Lubensky, D.,"Graph-based representation and techniques for NLU application development," Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on , vol.1, no., pp.I-288,I-291 vol.1, 6-10 April 2003.

[20] Hui Li; Wenju Liu; Honglei Han, "Graph Reduction Algorithm for Hardware/Software Partitioning," Control, Automation and Systems Engineering (CASE), 2011 International Conference on , vol., no., pp.1,4, 30-31 July 2011.

[21] Goldman, A., Ngoko, Y.," On Graph Reduction for QoS Prediction of Very Large Web Service Compositions," In: IEEE SCC, pp. 258–265 (2012).

[22] A Novel Technique for Call Graph Reduction for Bug Localization. International Journal of Computer Applications 47(15):1-5, Published by Foundation of Computer Science, New York, USA, June 2012.

[23] Rinehart, M.; Dahleh, M.A., "A graph reduction for bounding the value of side information in shortest path optimization," American Control Conference (ACC), 2010, vol., no., pp.4078, 4083, June 30 2010-July 2 2010

تقليل الرسوم الدلالية المستخدمة فى التلخيص التجريدى للنصوص

مروة محمود[١] , إبراهيم فتحى[٢],مصطفى عارف[٣]

قسم علوم الحاسب
كلية الحاسبات و المعلومات
جامعة عين شمس –القاهرة- مصر
[1]marwa_mahmoud100@hotmail.com
[2]ibrahim.moawad@heic.eg
[3]mostafa.m.aref@gmail.com

الملخص— تلخيص النصوص هو واحد من أهم تطبيقات المعالجة الطبيعية للغة. فهو يساعد المستخدمين فى التعامل مع الكميات الكبيرة من المعلومات المتاحة عن طريق ضغط محتويات الوثائق و إستخراج أهم الحقائق و الموضوعات التى تحتويها. تلخيص النصوص يمكن ان يصنف حسب نوع التلخيص إلى إستخراجى و تجريدى. فالتلخيص الاستخراجى يشمل تحديد أهم المقاطع فى النص كما هى أما التلخيص التجريدى فيستخرج أهم أجزاء النص بصياغة جديدة. في هذه الورقة البحثية, نعرض العديد من الطرق و الاساليب لتقليل و تصغير الرسوم .هذه الأساليب لها تطبيقات مختلفة في مجالات مختلفة . هذه الورقة جزء من بحث لتطبيق تقنيات تصغير الرسوم الدلالية فى مجال التلخيص التجريدى للنصوص.

هنسة عين شمس

جمعية هندسة اللغة

2013-12-11

فـروع الإنسانيـــات الجـدد وحوسبـــة اللغـــة

د. نبيل علي

INSPIRE OR EXPIRE

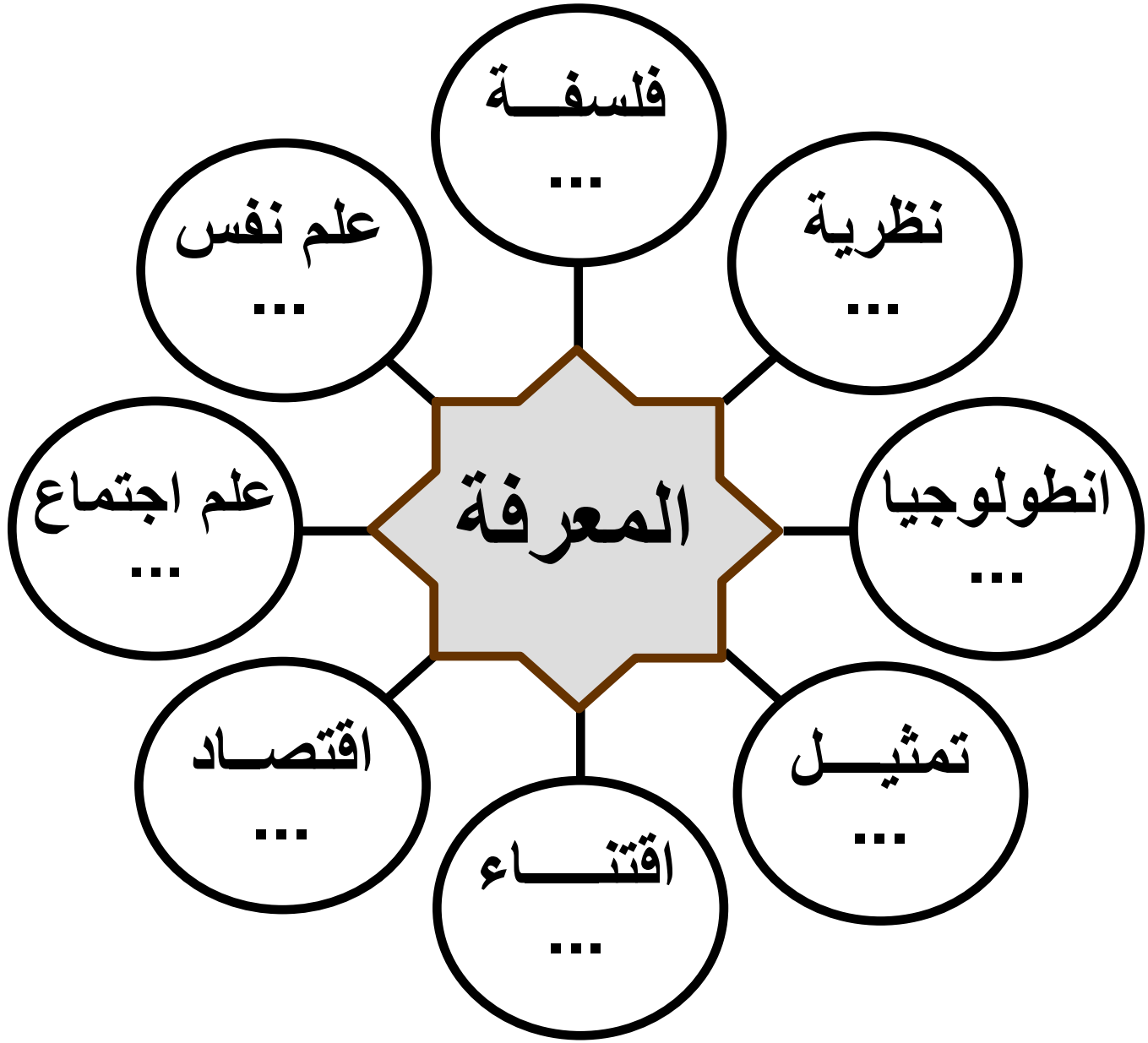
الإطار العام

1. الإنسانيات : النقلة النوعية
2. موسم الهجرة إلى الجمعي
3. حوار الشبكات
4. ثورة البيانات ومنهجيات حل المشكلات

علـــوم الانسانيـــات تجـــدد جلدهـــا ..!

علـــم الاجتمـــاع	المعرفي	SOCIOLOGY	COGNITIVE
علـــم النفـــــس	المعرفي	PSYCHOLOGY	COGNITIVE
علـــم اللغويـــات	المعرفي	LINGUISTICS	COGNITIVE
علـــم التاريـــخ	المعرفي	HISTORY	COGNITIVE
علـــم التربيـــة	المعرفي	PEDAGOGY	COGNITIVE
علـــم الجمـــال	المعرفي	AESTHETICS	COGNITIVE
علـــم الاقتصـــاد	المعرفي	ECONOMY	COGNITIVE

معـــــارف المعرفـــــــة



الذكاء الاصطناعي يلوذ بالفلسفة وعلوم والإنسانيات

- الذكاء الاصطناعي يستعصي على التطور من خلال تراكم التحسينات المتدرجة INCREMENTAL
- من خلال الفلسفة وعلوم الإنسانيات سنكتشف كم هي ضيقة نظرتنا للذكاء الاصطناعي
- لن يقلل ذلك من جاذبية أهل حوسبة اللغة في سوق العمل بل على العكس سوف يعززها
- حلم أصل الذكاء الاصطناعي هو محاكاة ما يجرى داخل المخ البشري، وإن تعذر ذلك حاليا فعلينا أن اقتفاء تجلياته المحسوسة ومظاهر سلوكه المختلفة

TIME GO COLLECTIVE موسم الهجرة إلى الجمعي

• COLLECTIVE INTELLIGENCE	• الذكاء الجمعي
• COLLECTIVE FILTERING	• الترشيح الجمعي
• COLLABORATIVE LEARNING	• التعليم التعاوني
• COLLABORATIVE KNOWLEDGE GENERATION	• توليد المعرفة تعاونيا
• COLLABORATIVE PROGRAMMING (OPEN SOURCE)	• تطوير البرامج تعاونيا
• CROWD SOURCING	• احتشاد المصادر
• SOCIAL SEARCH ENGINE	• محرك البحث الاجتماعي
• COLLABORATIVE CONSUMPTION	• الاستهلاك التعاوني
• PARTICIPATORY PLANNING	• التخطيط التشاركي

ما السر وراء كون الكثير أشد فطنة من القليل

WAY THE MANY IS SMARTER THAN THE FEW

حكمة الاحتشاد **CROWD WISDOM**

الاحتشاد من الغوغائية إلى الحكمة

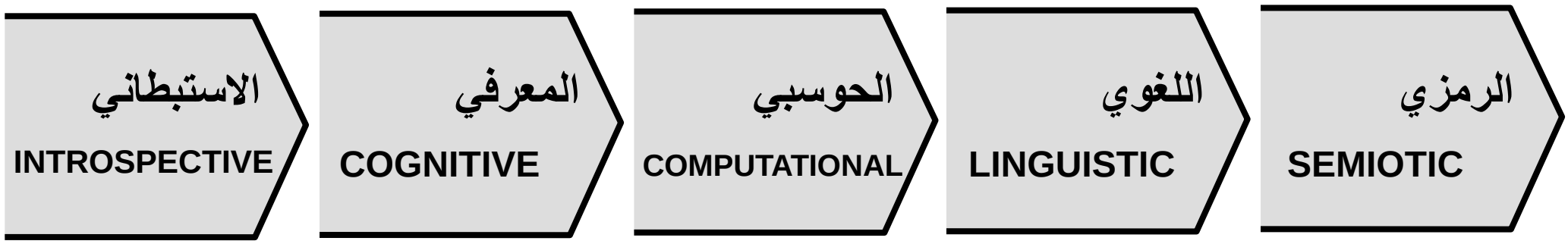
دعنا ننفض الضوضاء عن الظاهر المخادع الزائغ للكشف عن النظام الكامن في جوفه

التعلم العميق هو أداتنا للسيطرة على العشوائية السطحية للبيانات للكشف عما يعتملوا في جوفها من علاقات

ترديدات ثنائية الفردي والجمعي

النفسي PSYCHOLOGICAL	الاجتماعي SOCIOLOGICAL
استظهار EXTERNALIZE	استبطان INTERNALIZE
السردي NARRATIVE	الظاهراني PHENOMENOLOGICAL
الذاتي SUBJECTIVE	الموضوعي OBJECTIVE
الميكرو MICRO	الماكرو MACRO

سلسلة من النقلات النوعية



SEARCH ENGINES TO WHERE? محركات البحث إلى أين؟

مدخل البحث SEARCH ENTRY	لغويـــــا LINGUISTICS	حوسبة اللغـــــة LANGUAGE COMPUTATION
كلمات مفتاحية KEYWORDS	الصـــــرف MORPHOLOGY	معالجة الصرف آليا MORPHOLOGICAL PROCESSING
نصـــــي TEXTUAL	التركيـــب SYNTAX	الإعراب الآلـــــي AUTOMATIC PARSING
مفهومـــي CONCEPTUAL	الدلالـــــة SEMANTICS	الفهم الاتوماتي (ضحل/ عميق) AUT. UNDERSTANDING (SHALLOWLY/ IN-DEPTH)
اجتماعـــي SOCIAL	البرجماتيـــة PRAGMATICS	هندسة التخاطـــب CONVERSATIONAL ENGINEERING

حـــــوار الشبكات

NETWORK DIALOGUE

الشبكات الاجتماعيــة	الشبكات الأعصابية
SOCIAL NETWORK	NEURAL NETWORK

NETWORK DIALOGUE

حـــــوار الشبكات

الشبكات الأعصابية NEURAL NETWORK	الشبكات الدلاليـــــة SEMANTIC NET	الشبكات الاجتماعيــة SOCIAL NETWORK
انفجـــار البيانـــات DATA EXPLOSION	البنى النحوية المعجمية LEXICOS SYTACTICALS STRUCTURE	التفاعل الاجتماعي SOCIAL INTERACTION
التعلـــــم العميـــق DEEP LEARNING	السمـــات الدلاليـــة SEMANTIC FEATURES	علاقات التواصل البينية INTER- CONNECTIVITY

لا تفوق قدرة الإنسان على حل المشكلات إلا قدرته على خلق مشكلات جديدة

كورت جودل: مبدأ عدم الاكتمال

حل المشكلة معروف مسبقا

PROGRAM	=	ALGORITHM	+	DATA

حل المشكلة غير معروف

DEEP LEARNING	=	•NEURAL NET •DATA MINING	+	PREVIOUS DATA

التنقيب فـــي مناجـــم البيانـــات

DATA MINING

في عصرنا الرقمي افعالنا ورغباتنا وميولنا مسجلة إلكترونيا

- توقع الجرائم أوقاتها وأماكنها
- وقوع الحوادث
- تقلبات الأسواق
- معدلات الاستهلاك
- أنماط الطلب
- احتمالات الإصابة بالأمراض
- بيانات الفلك ومواقع النجوم

بيانـات ← **DM** → استخلصات

QUANTITATIVE METHODS

MATH. EQUATIONS

NEURAL NET

REGRESSION

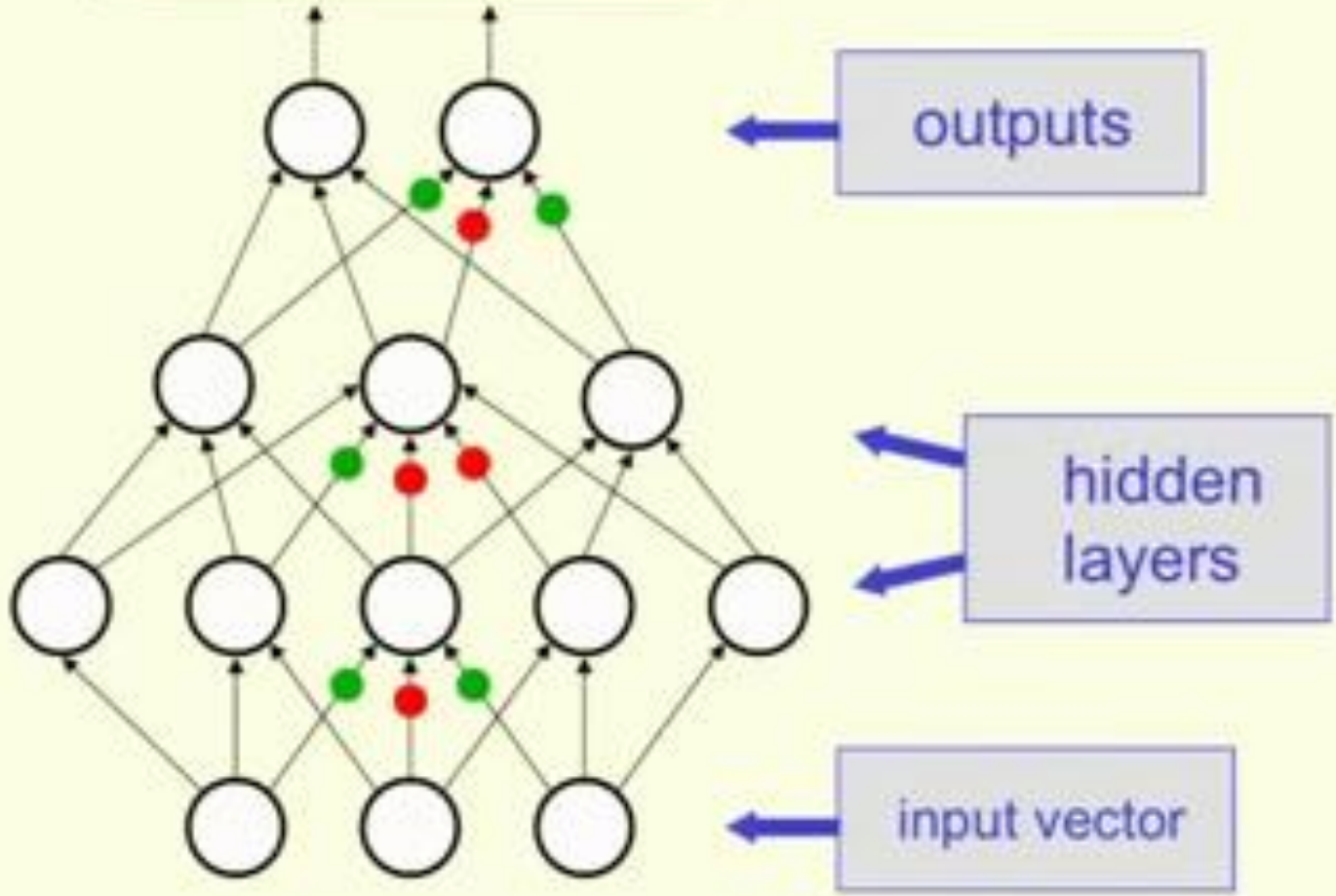
CLUSTERING

...

Deep neural networks (~1985)

Back-propagate error signal to get derivatives for learning

Compare outputs with correct answer to get error signal



طرق إيجاد حلول المشكلات

PROBLEM	SOLUTION	METHODOLOGY
KNOWN	KNOWN BY HUMANS	EXPERT SYSTEMS
KNOWN	AUTOMATIC	ALGORITHMIC / STATISTICAL
UNKNOWN	UNKNOWN	DEEP LEARNING DATA INTENSIVE GENERIC SOLUTIONS

PROBLEM SOLVING → PROBLEM INDEPENDENT

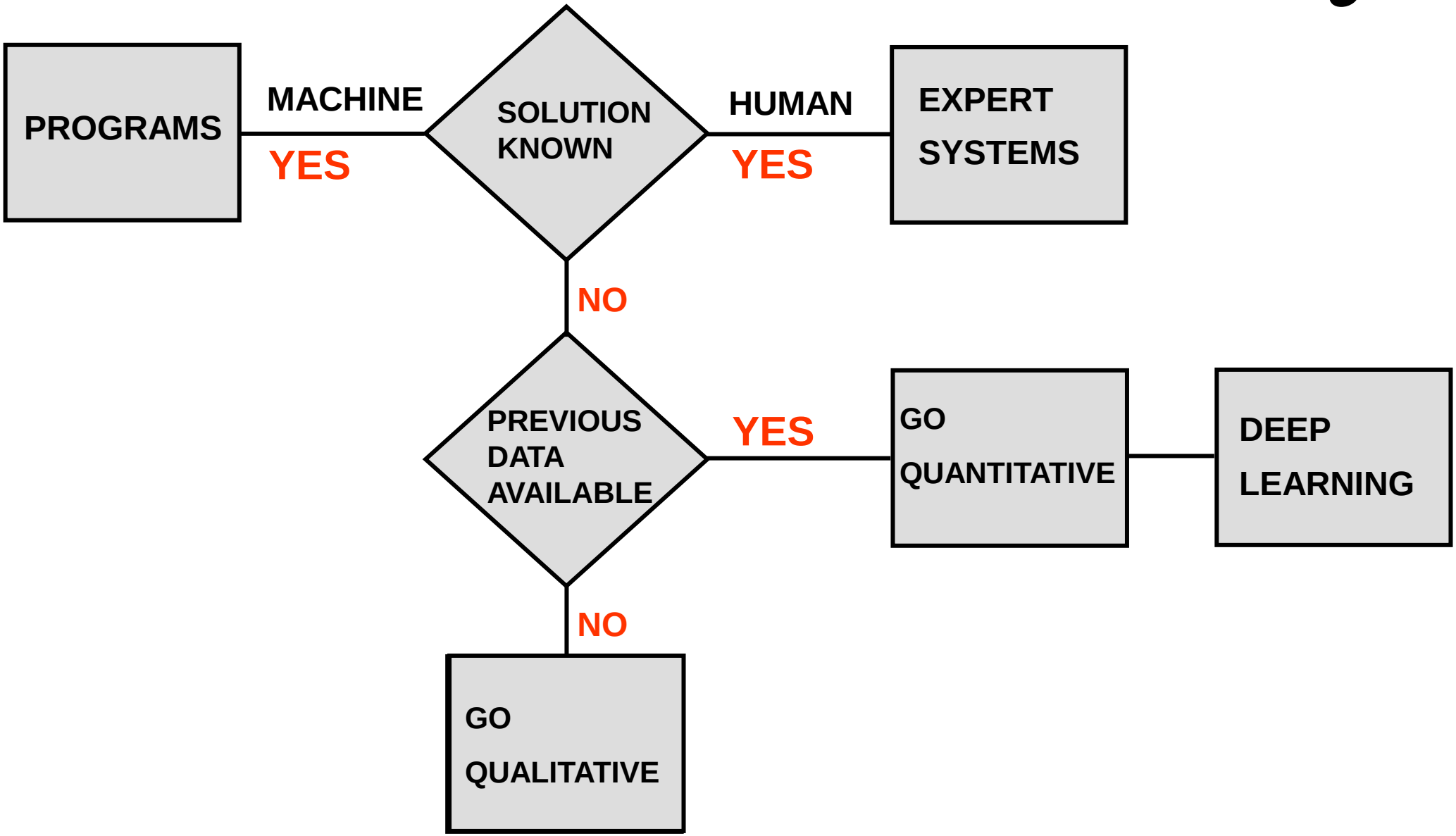
وداعا للتخصص البغيض

يا أهل حوسبة اللغة فلتسهموا في محاربة التخصص المنغلق

أداتكم البيانات وسلاحكم اللغة وزادكم الذي لا ينضب هو معرفة الإنسانيات

PROBLEM SOLVING

حـــل المشكـــلات



لا حلول مع اليأس ولا يأس مع الحلول

QUALITATIVE METHODS

- **NARRATIVE**
- **PHEMONELOGICAL**
- **GROUND THEORY**

Integrating Genetic Algorithms(GAs) with Conditional Random Fields(CRFs) to build A semi-supervised ANER system

Presented By:Noha Ahmd
Supervised by:
Prof Dr Aly Aly Fahmy
Prof Dr Ali Farghaly

Agenda

- Introduction
- Problem Statement
- Objective
- Motivation
- Background and Literature review
- Proposed solution
- Results
- Conclusion and future work

Introduction

- Named Entity Recognition and classification (NERC) is the process, by which proper names are identified and classified in unstructured texts and then classifying them into predefined classes such as person names, location, organization, and other named entity types.

Problem Statement

- Given a sequence of tokens in unstructured text ;
- **Example:**

لاسكندريه لكي يحضر مؤتمرا عن معالجة اللغات الطبيعيه

والذي عقد بمكتبة الاسكندريه

Problem Statement(Cont')

- Can we build a system that could detect
  - Person names
  - Organizations
  - Location
- In unstructured **Arabic text** and assign the right label to each of them, given a **small amount** of available labeled Arabic Data??

Objective

- the objective of this research is to :
  - Build An accurate ANER System using a small amount of supervision in order to recognize only three types of named Entities :
    - Person
    - Location
    - Organization

  In Arabic unstructured text.

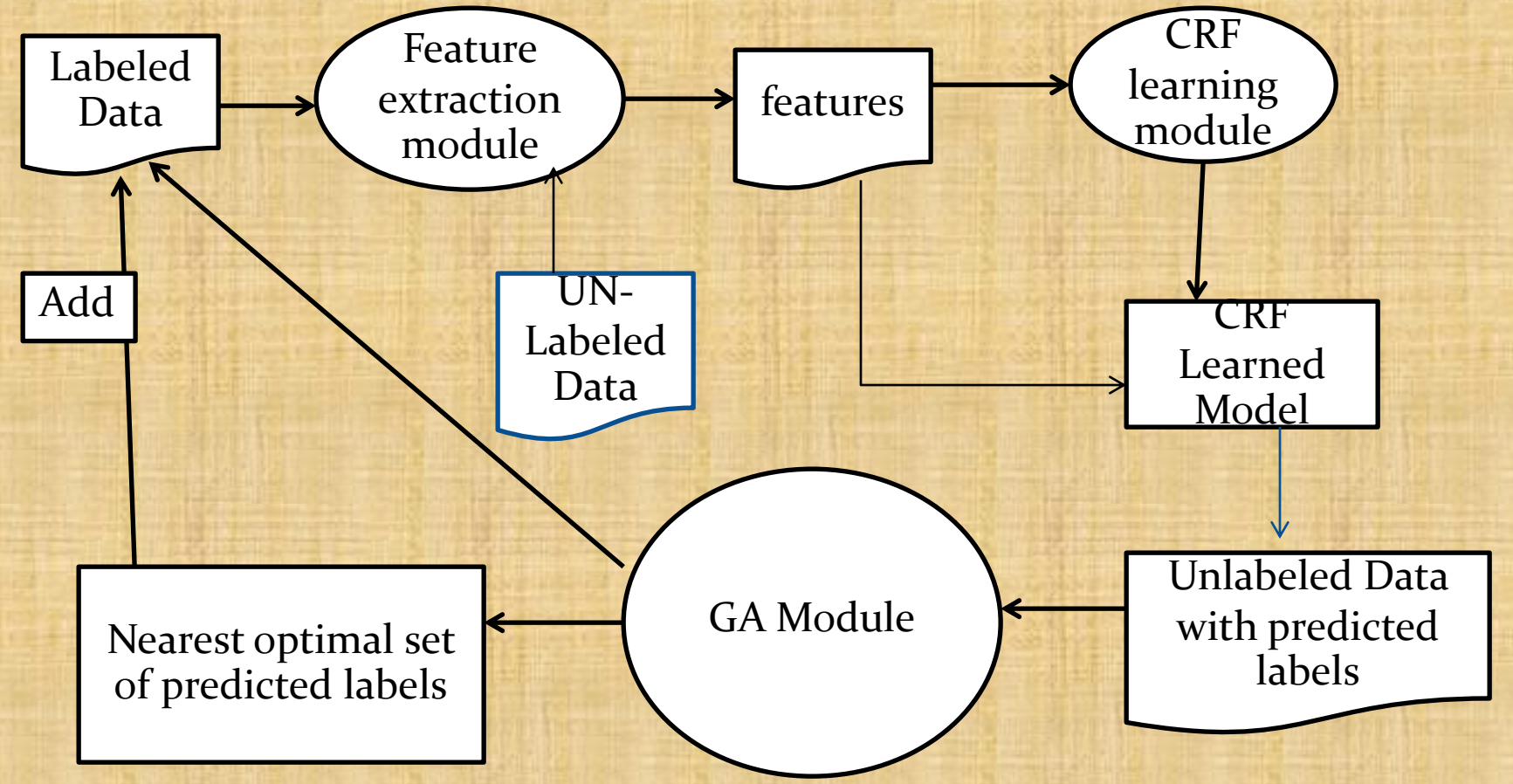
Motivation

- NER is considered an essential sub-task of many NLP
- Lack of accurate Arabic labeled data.
- Utilize unlabeled data that are exist with large amounts anywhere.

Proposed Solution

- Integrate between CRF And GA to build A semi-supervised learning ANER Systems

General View of SSL ANERC System


Labeled Data
Feature extraction module
features
CRF learning module
Add
UN-Labeled Data
CRF Learned Model
Nearest optimal set of predicted labels
GA Module
Unlabeled Data with predicted labels

Motivation to this solution

- The combination of classification methods may enhance the accuracy of the system .
- GA support the results of CRF by Selecting the optimum sequence of predicted labels produced by CRF .
- Many researches use CRF with GA in the feature selection process and achieve very good results

Motivation to this solution(cont')

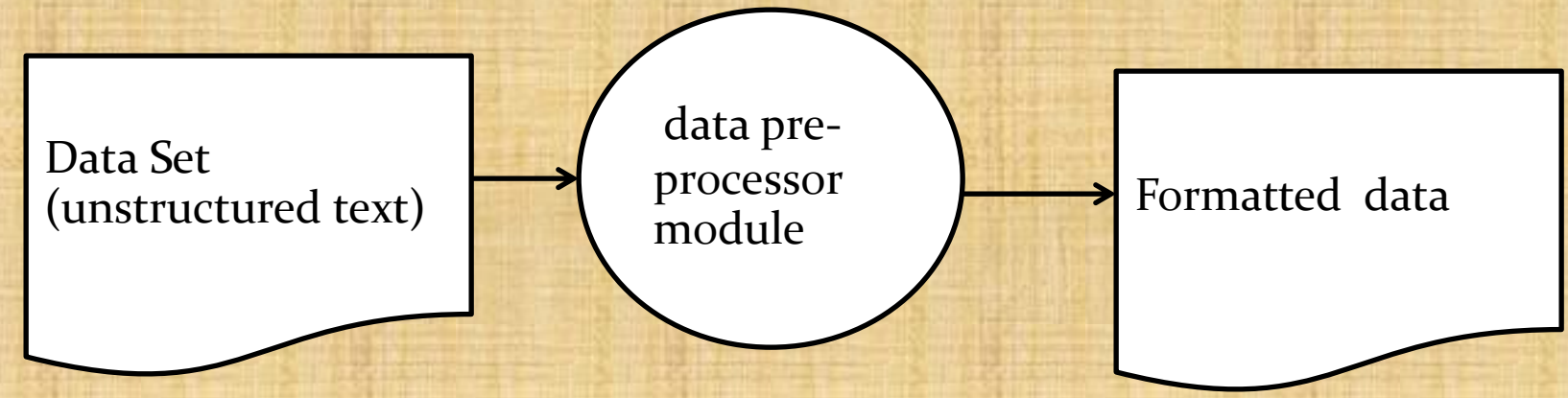
- No other researches have used this combination in semi-supervised learning in ANER field.
- The rational of using only the three types is because the main objective is to try the hybrid algorithm CRF and GA on the basic three types if it achieves good results it will be expand to cover more types of entities.

Proposed Solution Components

- Data pre-processing Module
- CRF Module
  - Training module
  - Testing module
- Pre-processor for genetic Algorithm
- Genetic Algorithm(GA)
- Evaluation Module

Data pre-processor Module

- This module is responsible for preparing data to be used in training and testing modules



Data pre-processing(cont’)

{ وقال كوفي أنان إن هذه الخطة يجب أن تشكل القاعدة السياسية لأي وقف لإطلاق النار }

{ وقال
كوفي
أنان
إن
هذه
الخطة
يجب
أن
تشكل
القاعدة
الأساسيه
لأي
وقف
لإطلاق
النار

Data pre-processing module

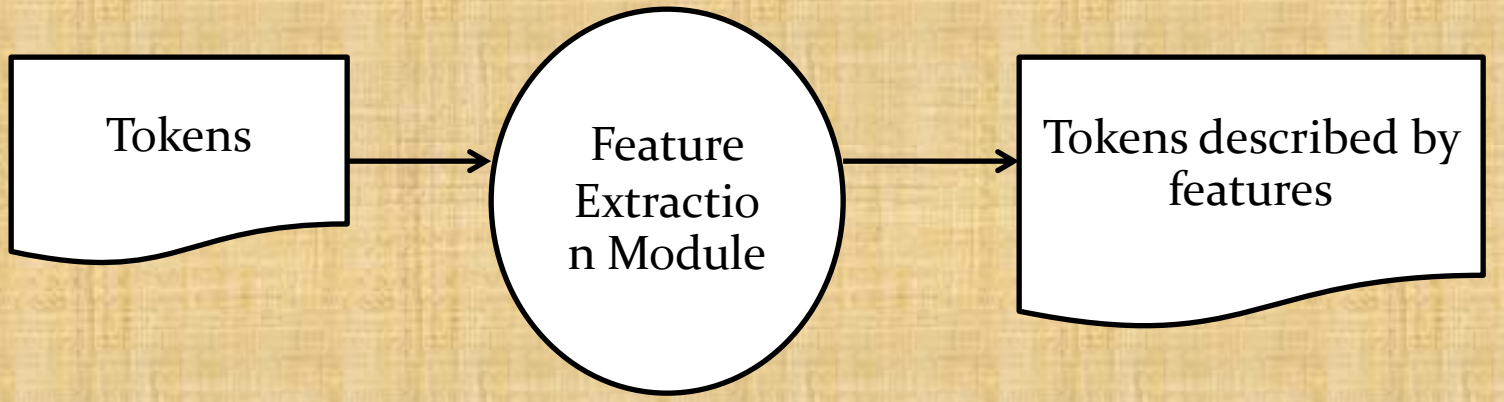
- Data cleaning
  - Data set contains some useless special characters, these special characters removed to make data more reasonable.
  - Such as :
  - “”
  -)
  - +,-,.,etc

Data pre-processing module

- Unifying the variation in spelling
  - The variation in spelling in data set reduces the accuracy of the system because the same word is seen as two different words.
  - Examples of variations:
    - أستراليا,استراليا
    - امريكا,اميركا
    - جرام,غرام
    - اسرائيل,إسرائيل
    - فاطمه,فاطمة

Data pre-processing module

- Feature Extraction Module
  - Data set itself can't be used to learn the computer, these data must be described by characteristics or attributes these characteristics or attributes called features.



Data pre-processing module

- Template File
  - While preparing training file ,A template file also is prepared
  - **Template File**

This file shows which features re used and how they are used .

Training Data

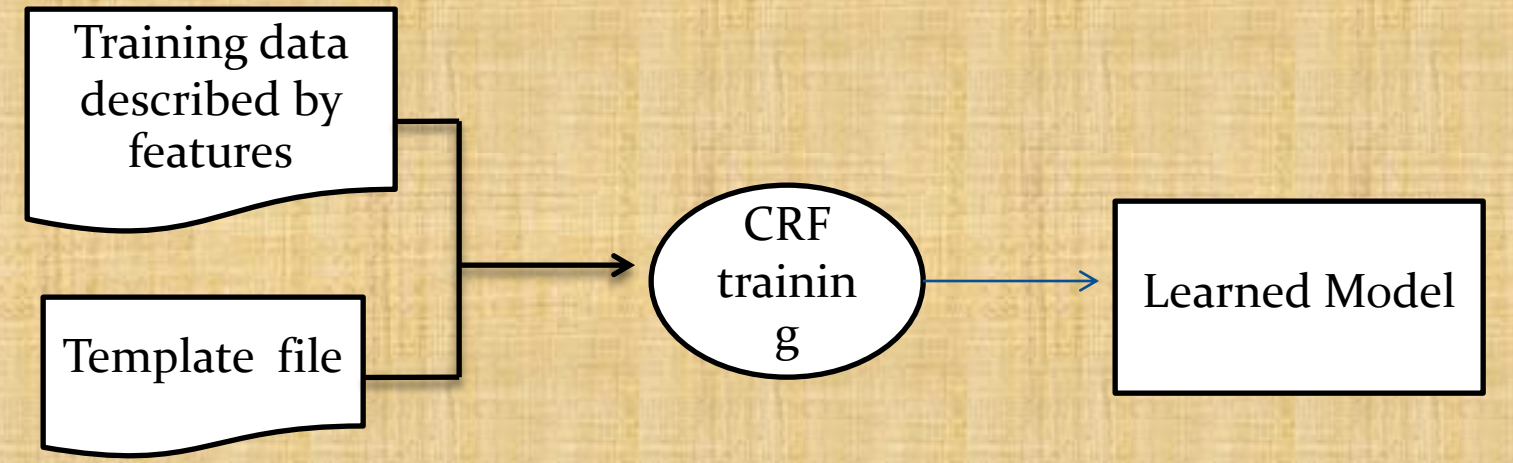
```
token person_gaz org_gz loc_gaz pers_Ind Org_ind loc_ind
اعلن false false false true false false O
اتحاد false false false false true false B-ORG
صناعة false false false false false false I-ORG
السيارات false false false false false false I-ORG
في false false false false false false O
المانيا false false true false false false B-LOC
امس false false false false false false O
الاول false false false false false false O
ان false false false false false false O
شركات false false false false true false O
صناعة false false false false false false O
السيارات false false false false false false O
في false false false false false false O
المانيا false false true false false false B-LOC
تواجه false false false false false false O
عاما false false false false false false O
صعبا false false false false false false O
في false false false false false false O
ظل false false false false false false O
ركود false false false false false false O
السوق false false true false true false O
الداخلية false false false false false false O
والصادرات false false false false false false O
وهي false false false false false false O
تسعي false false false false false false O
لان false false false false false false O
يبلغ false false false false false false O
الانتاج false false false false false false O
حوالي false false false false false false O
```

Template File

```
1  # Unigram
2  U00:%x[0,1]
3  U01:%x[0,0]
4  U02:%x[0,2]
5  U03:%x[0,3]
6  U04:%x[-1,4]/%x[0,1]
7  U05:%x[-2,4]
8  U06:%x[-3,4]
9  U07:%x[0,5]
10 U08:%x[-2,5]
11 U09:%x[-3,5]
12 U10:%x[-1,6]/%x[0,3]
13 U11:%x[-2,6]
14 U12:%x[-3,6]
15 u13:%x[-1,5]/%x[0,2]
16 U14:%x[-2,4]/%x[-1,4]/%x[0,1]
17 U15:%x[-2,5]/%x[-1,5]/%x[0,2]
18 U16:%x[-2,6]/%x[-1,6]/%x[0,3]
19
20
21 # Bigram
22 B
23
```

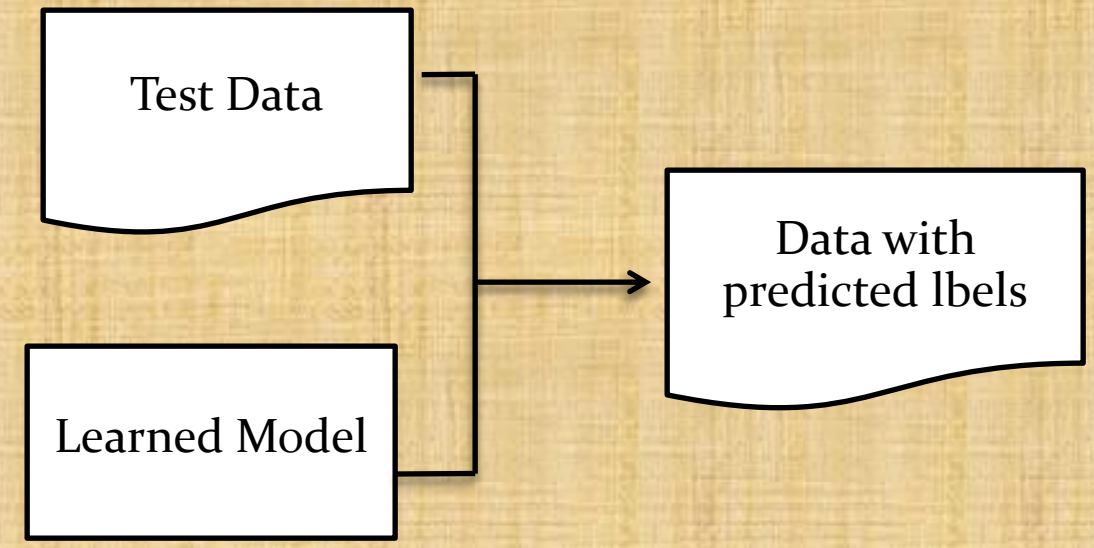
CRF Module

- CRF training module



CRF Module

- CRF Test



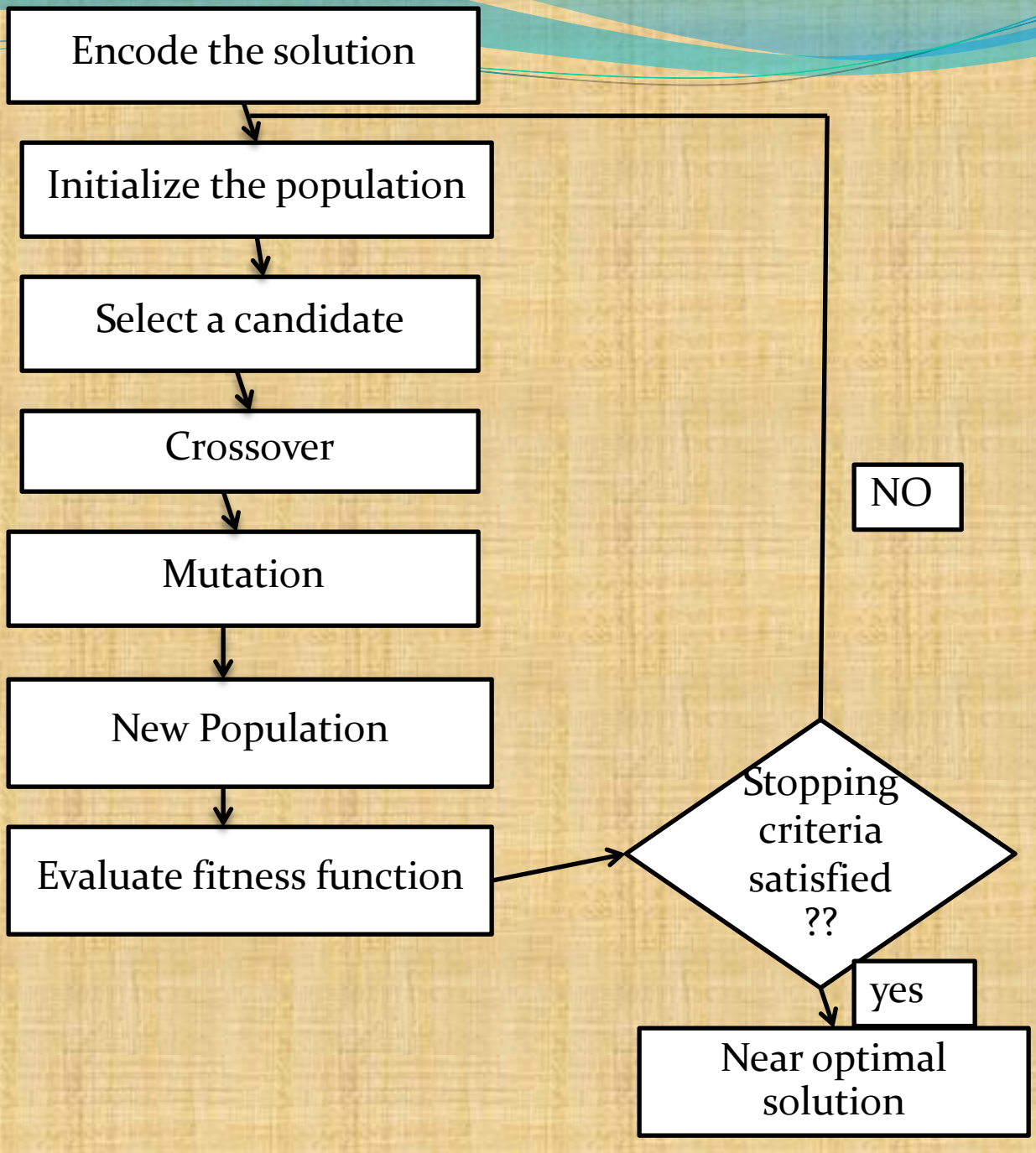
Tokens with Predicted labels

```
# 0.107630
وقال false false false true false false 0 O/0.995447 B-LOC/0.001635 B-ORG/0.000303 B-PERS/0.000999 I-LOC/0.000493 I-ORG/0.0
رئيس false false false false true true 0 O/0.921791 B-LOC/0.003444 B-ORG/0.065945 B-PERS/0.008323 I-LOC/0.000027 I-ORG/0.0
الاقتصاد false false false false true false 0 O/0.506515 B-LOC/0.022263 B-ORG/0.394094 B-PERS/0.020671 I-LOC/0.003468 I-ORG/0.0
برنيد false false false false false false B-PERS O/0.575552 B-LOC/0.001876 B-ORG/0.030599 B-PERS/0.028335 I-LOC/0.001446 I-ORG
جوتشولك false false false false false false I-PERS O/0.779117 B-LOC/0.002214 B-ORG/0.011409 B-PERS/0.053176 I-LOC/0.000934 I-ORG
عند false false false false false false 0 O/0.854442 B-LOC/0.021258 B-ORG/0.010350 B-PERS/0.013716 I-LOC/0.001775 I-ORG/0.06493
اعلان false false false false false false 0 O/0.947196 B-LOC/0.012703 B-ORG/0.005743 B-PERS/0.004216 I-LOC/0.002257 I-ORG/0.0
آخر false false false false false false 0 O/0.964574 B-LOC/0.013420 B-ORG/0.005740 B-PERS/0.004594 I-LOC/0.001407 I-ORG/0.00660
تقرير false false false false false false 0 O/0.967577 B-LOC/0.013481 B-ORG/0.005622 B-PERS/0.004617 I-LOC/0.001397 I-ORG/0.0
سنوي false false false false false false 0 O/0.969008 B-LOC/0.013409 B-ORG/0.005164 B-PERS/0.004415 I-LOC/0.001390 I-ORG/0.0
للاقتصاد false false false false false false 0 O/0.975033 B-LOC/0.011900 B-ORG/0.003170 B-PERS/0.003073 I-LOC/0.001294 I-ORG/0.0
ان false false false false false false 0 O/0.995089 B-LOC/0.002174 B-ORG/0.000903 B-PERS/0.000716 I-LOC/0.000256 I-ORG/0.00056
مستقبل false false false false false false 0 O/0.966577 B-LOC/0.024320 B-ORG/0.004131 B-PERS/0.002763 I-LOC/0.000593 I-ORG/0.0
السوق false false true false true false 0 O/0.408970 B-LOC/0.556862 B-ORG/0.025754 B-PERS/0.000346 I-LOC/0.006870 I-ORG/0.0
مازال false false false false false false 0 O/0.732558 B-LOC/0.007265 B-ORG/0.015560 B-PERS/0.003767 I-LOC/0.180432 I-ORG/0.0
يفتقر false false false false false false 0 O/0.928970 B-LOC/0.001634 B-ORG/0.013305 B-PERS/0.014921 I-LOC/0.005609 I-ORG/0.0
الى false false false false false false 0 O/0.910172 B-LOC/0.023388 B-ORG/0.012278 B-PERS/0.017153 I-LOC/0.001317 I-ORG/0.02600
الخطوط false false false false false false 0 O/0.940535 B-LOC/0.015730 B-ORG/0.010087 B-PERS/0.007878 I-LOC/0.002481 I-ORG/0.0
الواضحة false false false false false false 0 O/0.912935 B-LOC/0.033098 B-ORG/0.016571 B-PERS/0.009290 I-LOC/0.004301 I-ORG/0.0

# 0.120801
خطة false false false false false false 0 O/0.996872 B-LOC/0.000546 B-ORG/0.000203 B-PERS/0.001404 I-LOC/0.000195 I-ORG/0.00037
انان true false false false false false B-PERS B-PERS/0.986331 B-LOC/0.000917 B-ORG/0.000493 B-PERS/0.986331 I-LOC/0.000036 I
و true true true false true false 0 O/0.962014 B-LOC/0.005011 B-ORG/0.001028 B-PERS/0.001858 I-LOC/0.003886 I-ORG/0.00169
في false false false false false false 0 O/0.986865 B-LOC/0.003856 B-ORG/0.002735 B-PERS/0.001343 I-LOC/0.001773 I-ORG/0.00270
نيويورك false false true false false false B-LOC O/0.686455 B-LOC/0.271477 B-ORG/0.004223 B-PERS/0.022723 I-LOC/0.012029 I-ORG
دعا false false false true false false 0 O/0.893388 B-LOC/0.025634 B-ORG/0.006877 B-PERS/0.008899 I-LOC/0.043699 I-ORG/0.00849
الأمين false false false false false false 0 O/0.986094 B-LOC/0.001017 B-ORG/0.004198 B-PERS/0.003172 I-LOC/0.000724 I-ORG/0.0
العام false false false false false false 0 O/0.933398 B-LOC/0.008310 B-ORG/0.026071 B-PERS/0.022313 I-LOC/0.000285 I-ORG/0.0
```

Genetic Algorithm

- A GA is developed to enhance the results come from the CRF.


Encode the solution
Initialize the population
Select a candidate
Crossover
Mutation
New Population
Evaluate fitness function
Stopping criteria satisfied ??
NO
yes
Near optimal solution

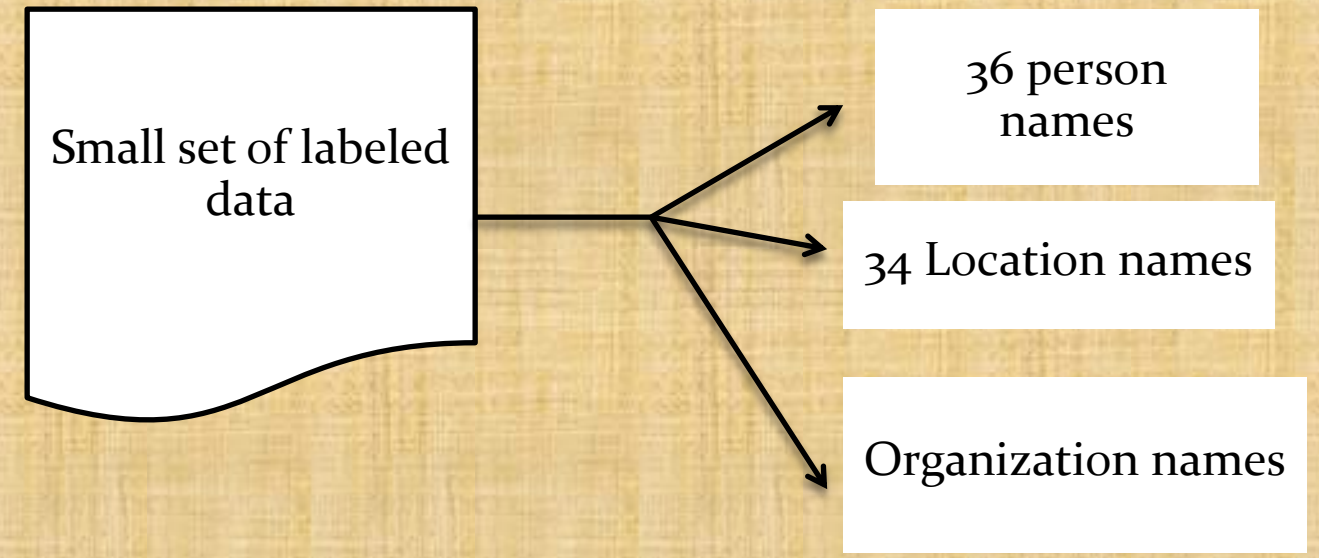
Encoding the solution

- This process aims to set the structure of the chromosome to represent a feasible solution.
- what is a solution of the GA module in this work?
- GA module is applied to set of unlabelled data with predicted labels come from CRF testing.

Encoding the solution(cont’)

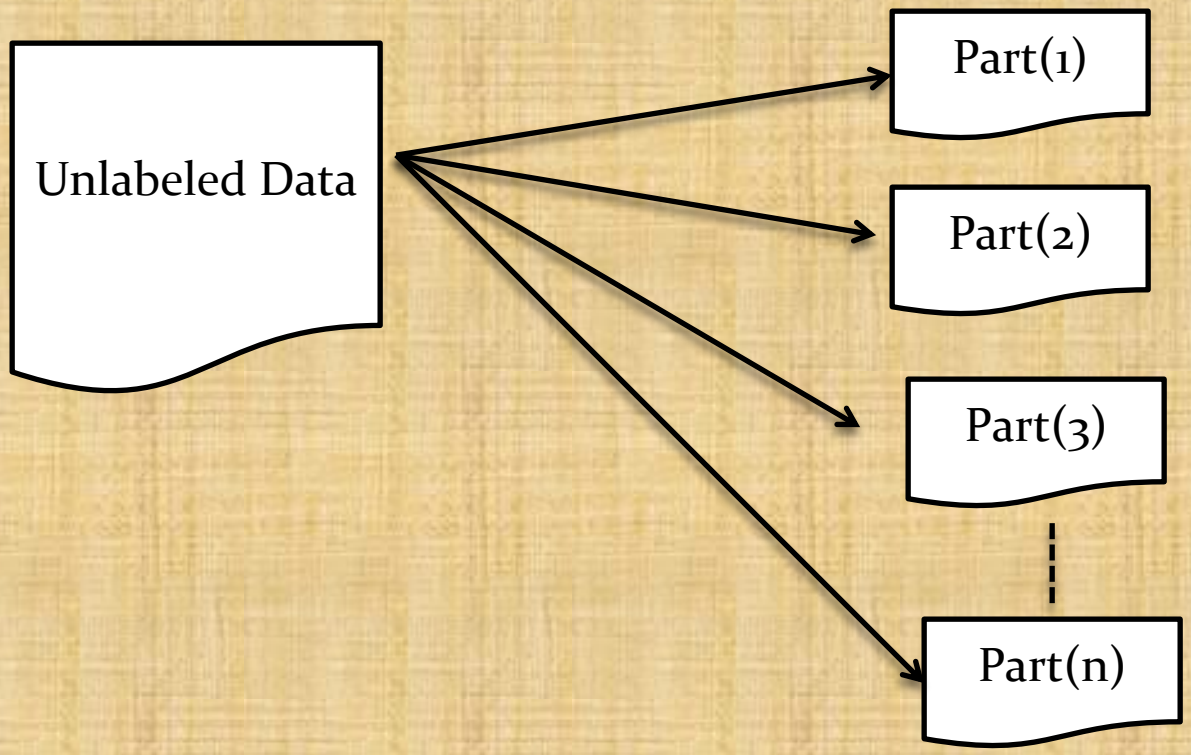
- As mentioned before ,the objective of GA is to add unlabelled data with predicted labels to the labeled training data to maximize its size and therefore enhance the accuracy of the system

Initial set of labeled data(Seeds)

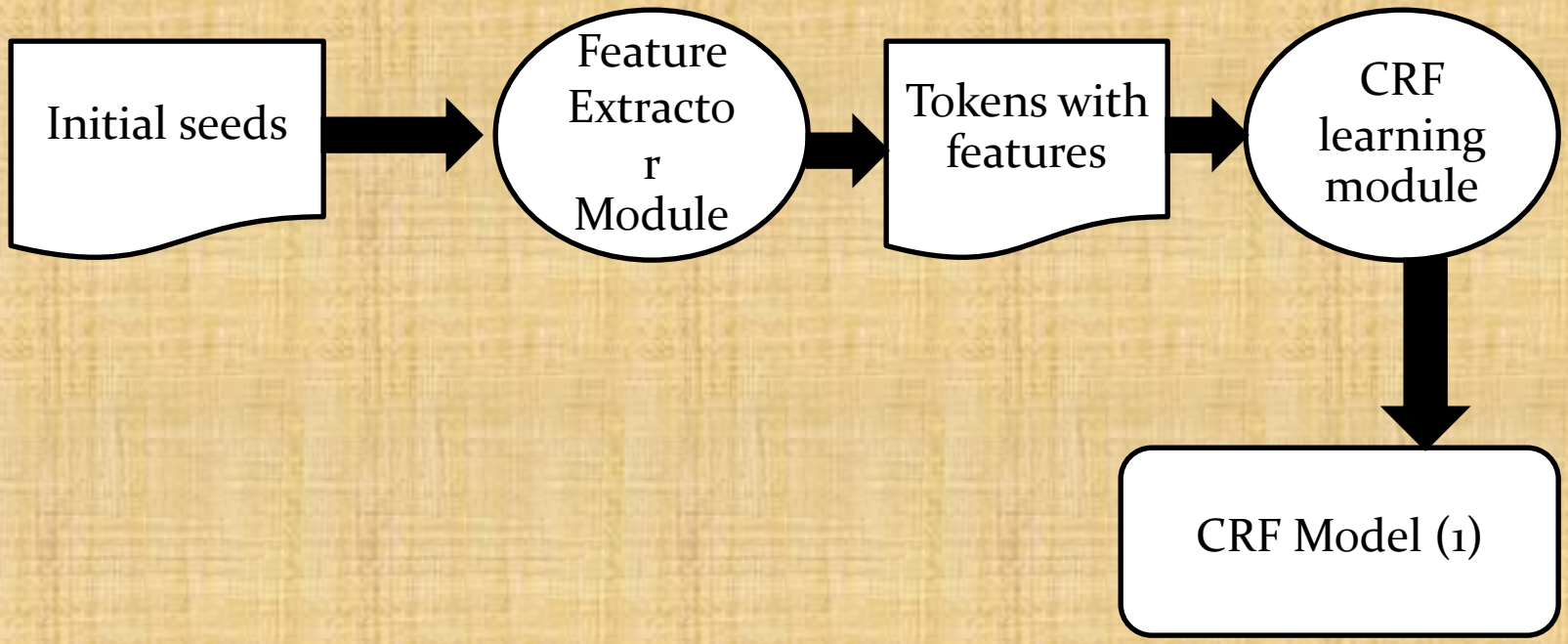


Unlabeled Data

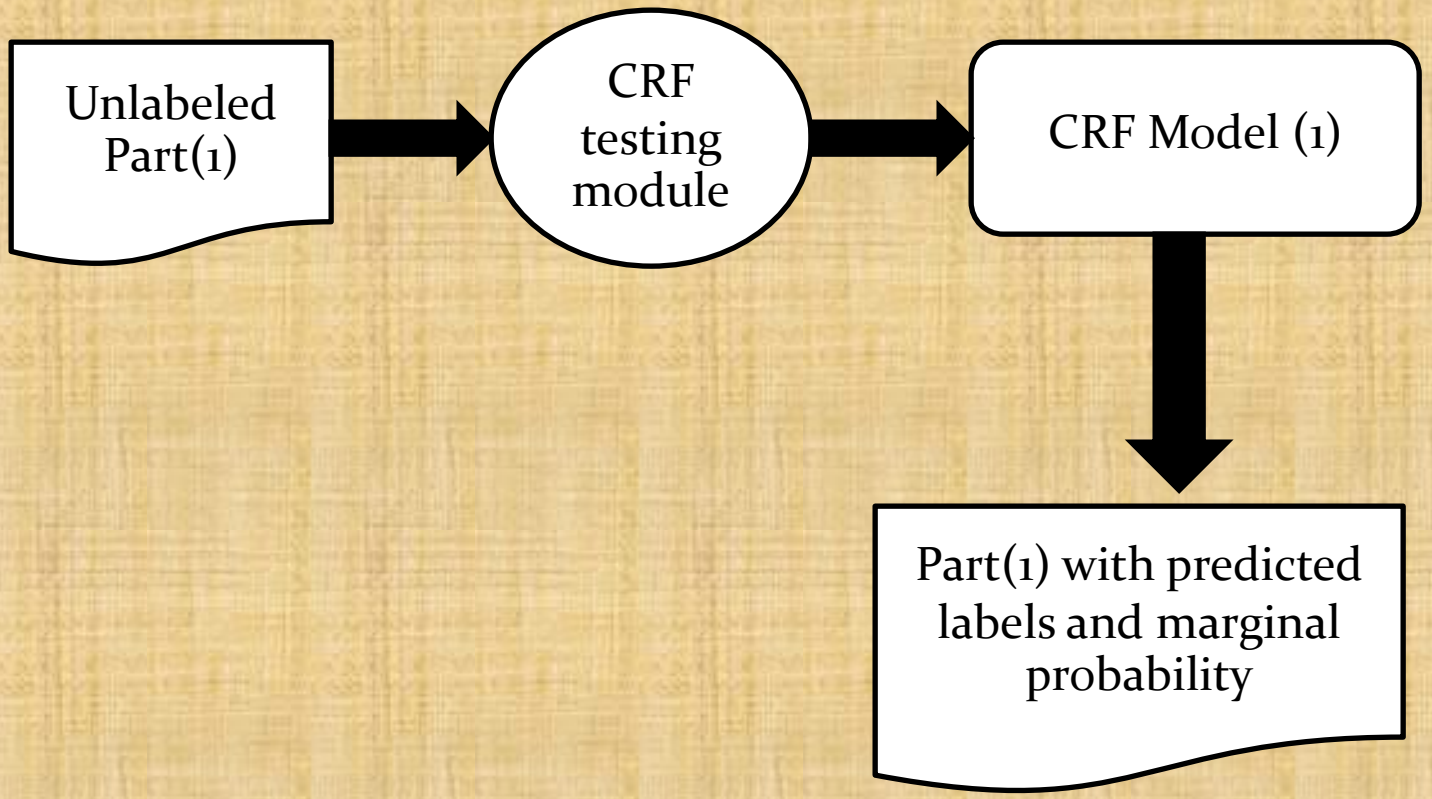
- Unlabeled data is partitioned into some parts each part is about 400 tokens

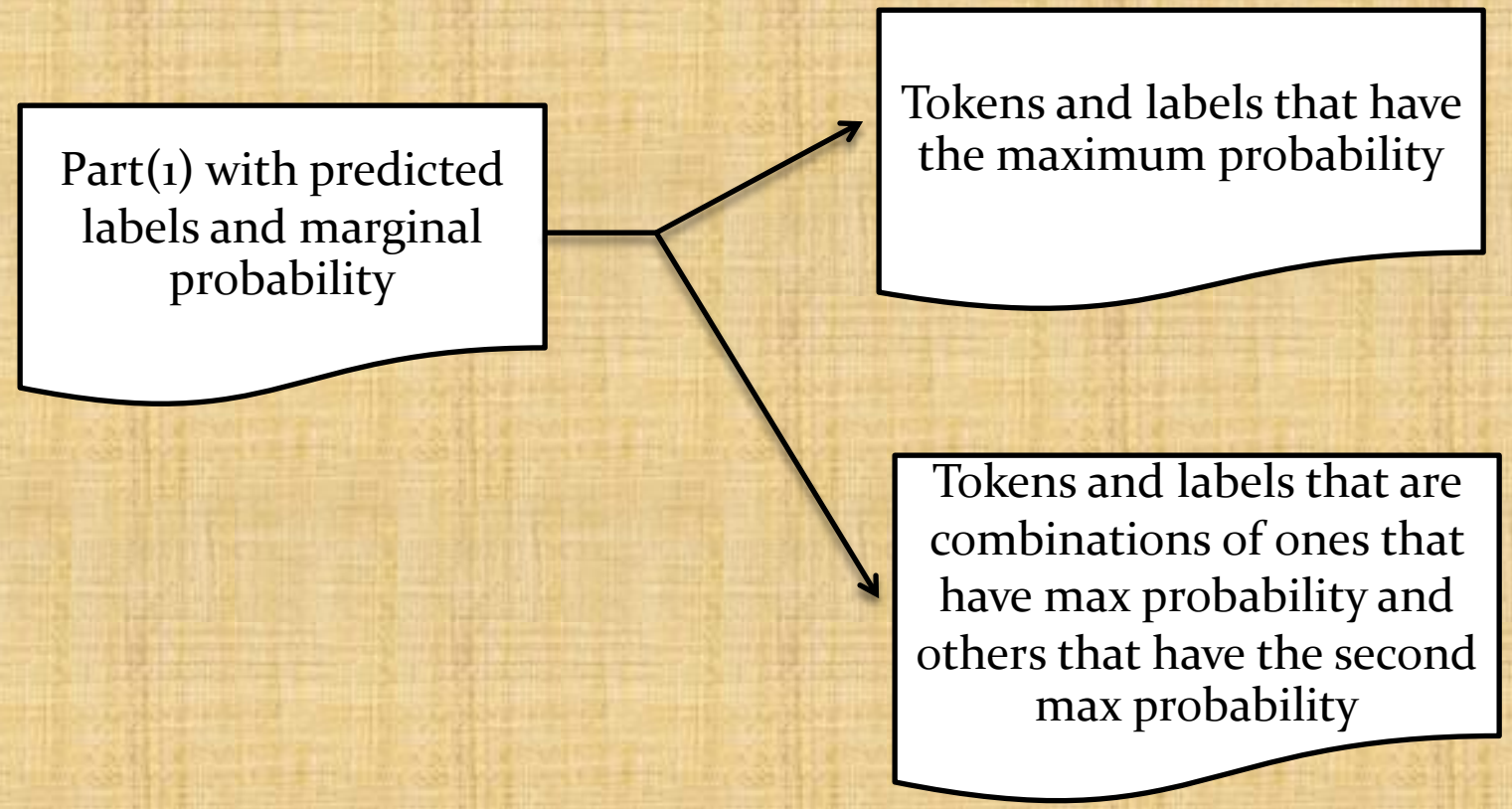


Iterations and unlabeled data addition



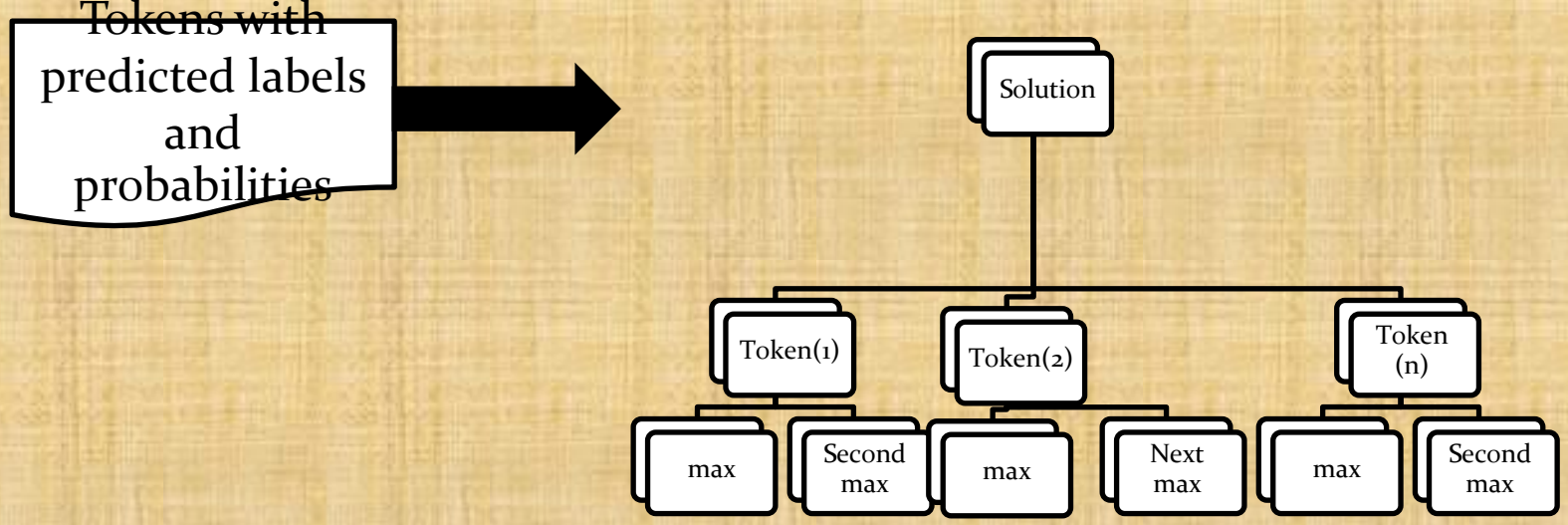
Iterations and unlabeled data addition




Part(1) with predicted labels and marginal probability
Tokens and labels that have the maximum probability
Tokens and labels that are combinations of ones that have max probability and others that have the second max probability

Using only CRF

Using both CRF and GA



Max and second max Encoding

- Max  0
- Second Max  1

Chromosome Encoding

- The chromosome length is the length of the part of unlabeled data.
- Each gene in the chromosome represents a token in the unlabeled data.
- The index of the gene is the index of the token
- The value of the gene is either :

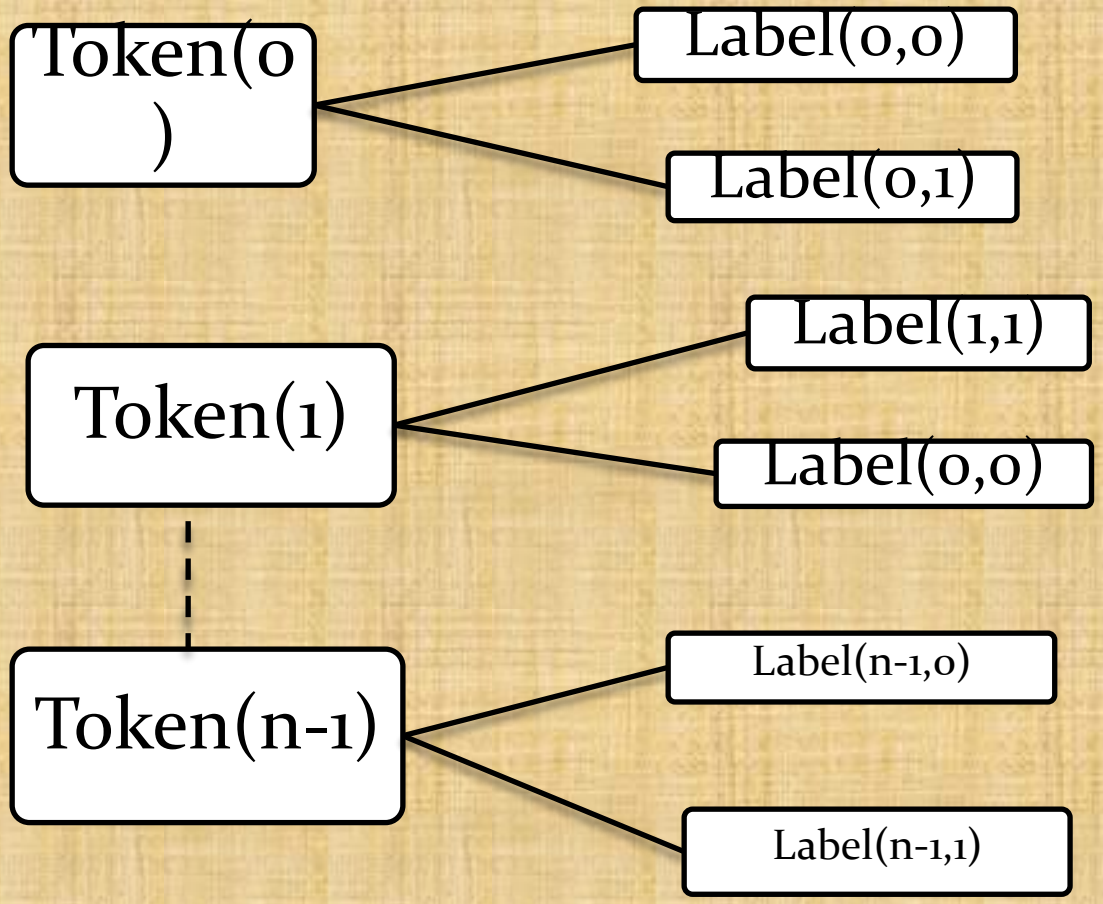
  0→the token takes the label with max probability

  1→the token takes the label with second next probability

Chromosome Encoding

0	1	2	3	4	5	6	7	....	Size-1
0	1	1	0	1	0	0	1	1	0

Example (part(1))



The problem

- Select the best labels given max and second max for the sequence of unlabeled tokens.
- Searching for this best sequence given a search space of all combinations of max and second max labels is very costly.

Using approximate search technique

- Genetic Algorithm(GA)
  - Create a population of 30 candidates
  - Set score to all chromosomes
  - Evaluate the fitness function
  - Do GA operators
    - Crossover
    - mutation

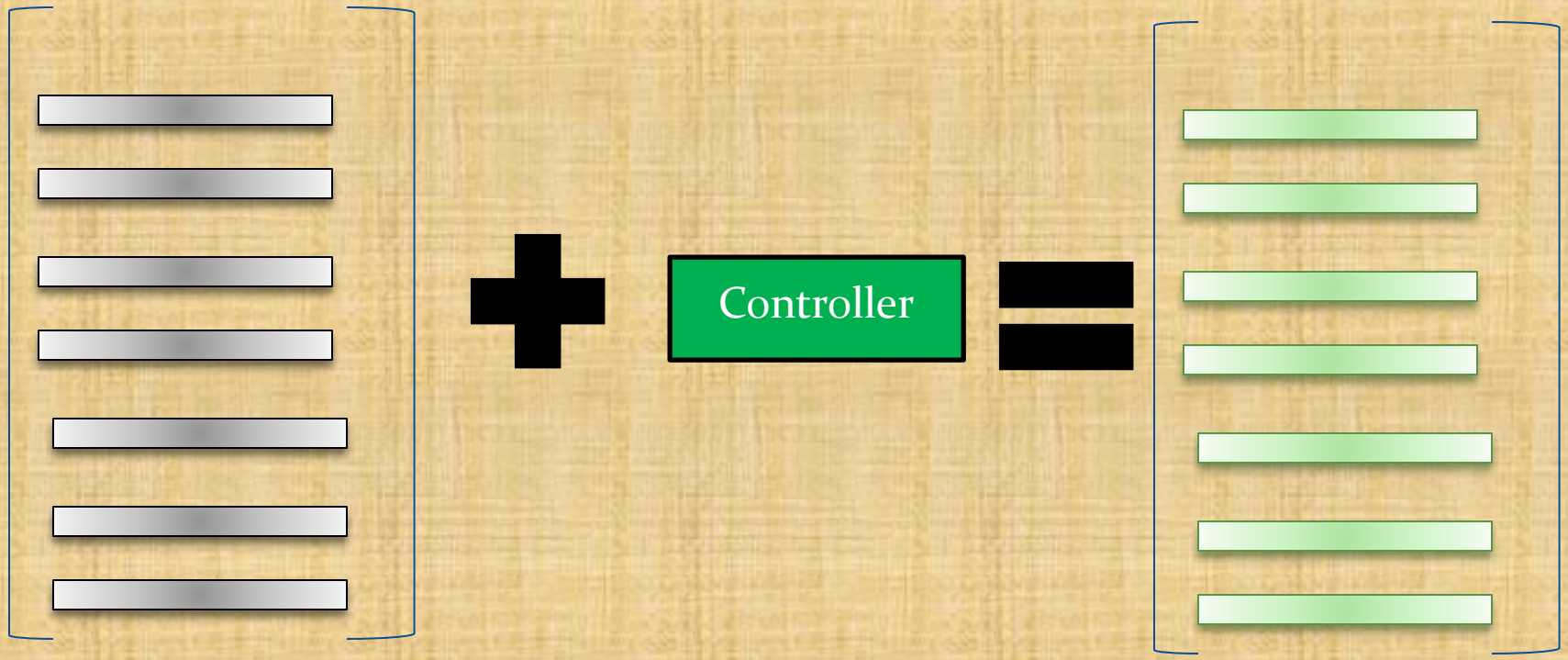
population

- The initial population is not fully random!!!
- The generation of the population is controlled

controller

- This chromosome is called controller .
- Genes have values one if only the difference between the max and second max is less than or equal to 0.2.
- All random generated candidates is and with this chromosome to prevent it to divert from the accurate results

Population and controler



Random population

Controlled population

GA operators

- Selection
  - Roulette wheel
- Crossover
  - single point crossover
- Mutation

Literature Review

- Arabic Named Entity Recognition Using CRF
- Arabic Named Entity Recognition Using Simplified Feature Set.
- Integrated Machine Learning Techniques For Arabic Named Entity Recognition

(1)Arabic Named Entity Recognition Using CRF

- Authors:
  - Yassin Ben Ajibaa
  - Paolo Rosso
- Year:
  - 2008
- Contribution:
  - Using CRF instead of Maximum Entropy (ME)in order to enhance their previous work which is developed using ME

Results using ME

	Precision	Recall	F-Measure
LOC	91.6	82.23	86.7
ORG	47.9	45.02	46.4
PERS	56.2	48.56	52.9
Overall	70.2	62.08	65.91

Results Using CRF

	Precision	Recall	Overall
LOC	93.03	86.14	89.74
ORG	84.23	53.94	65.76
PERS	80.41	67.42	73.35
Overall	86.90	57.83	79.21

observations

- It is clear from these results that CRF outperforms ME given the same feature set
- This is considered a proof that CRF achieves best results in Sequences problems like NER

(2)ANER Using Simplified Feature Set

- Authors:
  - Ahmed Abdelhameed
  - Kareem Darwish
- Year:2009 -2010

ANER Using Simplified Feature Set(cont')

- Contributions:
  - They have trained CRF on features that are primarily use character n-gram of leading and trailing letters in words and also word n-gram.
  - Their feature set helped to overcome some of the morphological and orthographic complexities of Arabic

ANER Using Simplified Feature Set(cont')

- Comparing their results in literature using Arabic specific features such as part of speech tagging on the same data set and same implementation of CRF
  - Although the results are lower by 2 F-Measure for locations
  - They outperformed the best results Benajiba has achieved overall

ANER Using Simplified Feature Set(Results)

	Precision	Recall	F-Measure
LOC	93%	83%	88%
ORG	84%	65%	74%
PERS	90%	75%	82%
Overall	89%	74%	81%

Hybrid Systems

- From single classifier to hybrid Systems
  - Integrated Machine Learning Techniques For Arabic Named Entity Recognition

Integrated Machine learning techniques for Arabic Named Entity Recognition

- Authors
  - Samier Abdelrahman
  - Mohammed Elarnaoty
  - Marwa Magdy
  - Aly Fahmy
- Year of Publication:
  - 2010

Contribution

- The solution is an integration approach between two machine learning techniques, namely:
  - bootstrapping semi-supervised pattern recognition
  - Conditional Random Fields (CRF)classifier as a supervised technique.
- The contributions are the exploit of pattern and word semantic fields as CRF features, the adventure of utilizing bootstrapping semi supervised pattern recognition technique in Arabic Language, and the integration success to improve the performance of its components.

Integrated Machine learning techniques for Arabic Named Entity Recognition

	precision	Recall	F-measure
LOC	96.05%	80.86%	87.80%
ORG	84.95%	60.02%	70.34%
PERS	89.20%	54.68%	67.80%
overall	90.06%	65.18%	75.31%

Our system results

- [1] Baseline
  - The model generated here is trained using a small set of labeled data that includes:
    - 36 person names
    - 34 location names
    - 28 organization names

    - This model is considered the main seeds for our semi-supervised model

Base line (supervised part)

	precision	Recall	F-measure
LOC	90.75%	75.78%	82. 59%
ORG	70%	43.64%	53.76%
PERS	39.42%	31.81%	35.20%
overall	69.39%	53.41%	57.18%

Part (1)

Using only CRF

	Precision	Recall	F-Measure
LOC	92.59%	79.78%	85.71%
ORG	80%	43.24%	56.14%
PERS	43.75%	31.81%	36.84%
OVERALL	72.11%	51.61%	59.65%

Using CRF and GA

	Precision	Recall	F-measure
LOC	93.75%	79.78	86.20
ORG	73.07	51.35	60.31
PERS	47.05	36.36	41.02
Overall	71.29%	55.83	62.51

Part (2)

Using only CRF

	precision	Recall	F-measure
LOC	94.93%	79.78%	86.70%
ORG	72.00%	48.64%	58.06%
PERS	40.00%	31. 81%	35.44%
overall	68.97%	53. 41%	60.06

Using CRF and GA

	Precision	Recall	F-measure
LOC	94.93%	79.78%	86.70%
ORG	77.27%	45.94%	57.62%
PERS	46.87%	34.09%	39.47%
OVERALL	73.02%	53.27%	61.26%

Part (3)

Using only CRF

	Precision	recall	F-measure
LOC	92.59%	79.78%	85. 71%
ORG	80%	43.24%	56.14%
PERS	44.11%	34.09%	38.46%
overall	72.23%	52.37%	60.10%

Using CRF and GA

	Precision	Recall	F-measure
LOC	93.75%	79.78%	86.20%
ORG	81.81%	48.64%	61.01%
PERS	44.73%	38.63%	41.46%
Overall	73.43%	55.68%	62.89%

Part (4)

Using only CRF

	Precision	Recall	F-measure
LOC	93.58%	77.65%	84.38%
ORG	81.81%	48.64%	61.01%
PERS	36.84%	31.81%	34.14%
overall	70.74%	52.7%	59.84%

Using CRF and GA

	Precision	Recall	F-measure
LOC	92.59%	79.78%	85.71%
ORG	80%	43.24%	56.14%
PERS	45.94%	38.63%	41.97%
OVERALL	72.84%	53.88%	61.27%

Part (5)

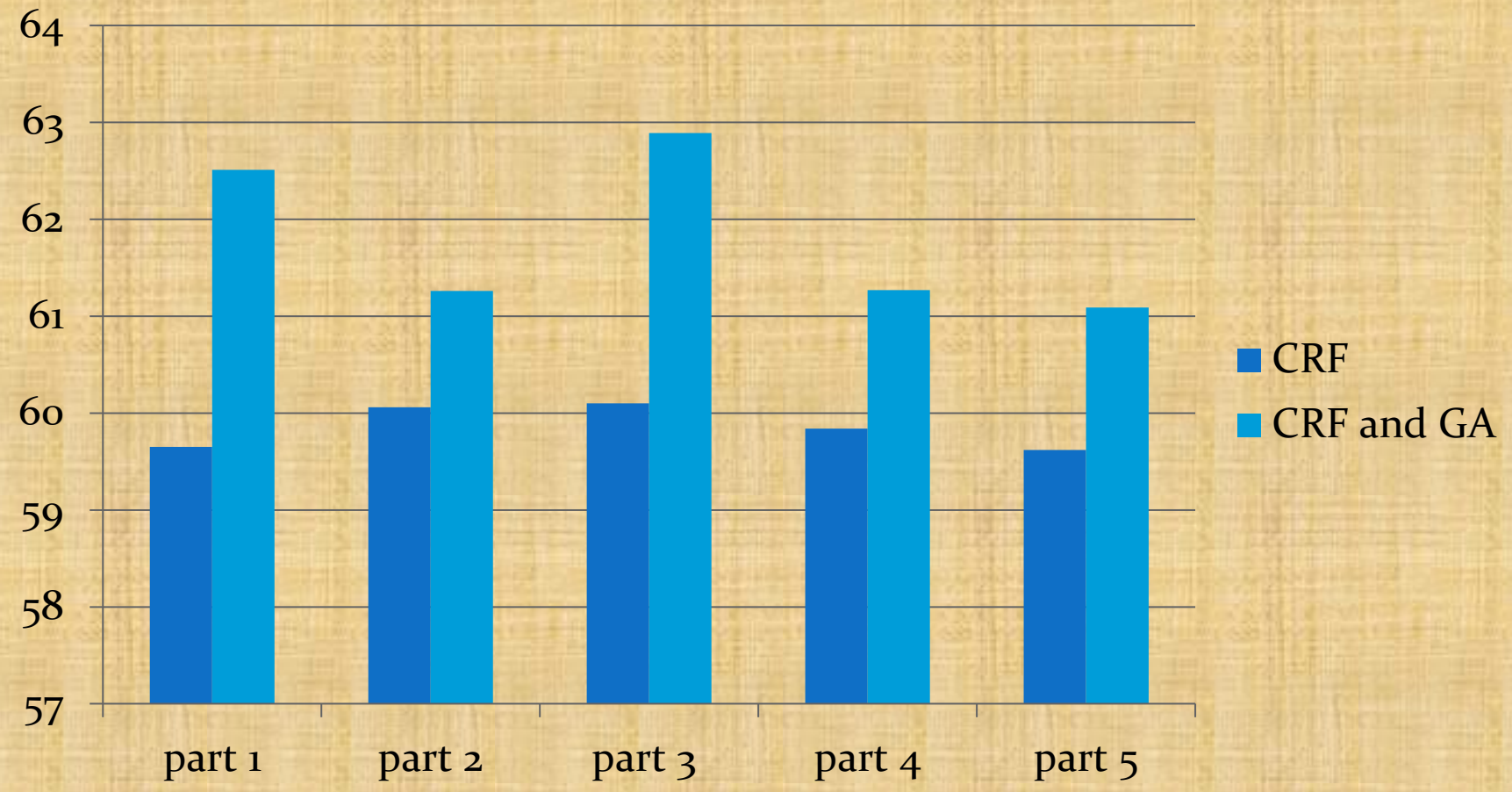
Using only CRF

	precision	recall	F-measure
LOC	92. 59%	79.78%	85.71%
ORG	80%	43.24%	56.14%
PERS	40.54%	34.09%	37.03
overall	71.04%	52.37%	59.62%

Using CRF and GA

	Precision	Recall	F-measure
LOC	93.75%	79.78%	86.20%
ORG	80.95%	45.94%	58.62%
PERS	44.11%	34.09%	38.46%
OVERALL	72.93%	53.27%	61.09%

Summary of results



conclusion

- the integration between GA algorithms and CRF outperforms Using The CRF only in all parts
- Not always adding new unlabeled data to the training data enhance the results


ANY
questions?


Thank
You!

أحكام تنافر صوتي الفعل الثلاثي المضعف

دراسة لغوية حاسوبية

أ. د. وفاء كامل فايد

كلية الآداب – جامعة القاهرة

تمهيد

- في بحوث سابقة دَرَسْتُ أثر تجاور صوتي الفعل الثلاثي المضعف على بابه الصرفي ، ورصدتُ عددا من القواعد التي تربط بين أصوات هذا الفعل واتجاهه إلى التصرف على باب صرفي بعينه.
- وهذا البحث استكمال للبحوث السابقة ، وبلورة لنتائجها باستخلاص القواعد التي توصلت إليها تلك البحوث ، ومحاولة ربطها في أسس عامة شاملة.

مقدمة

توصلت البحوث السابقة إلى أن العلاقة بين صوتي الفعل الثلاثي المضعف وبابه الصرفي تمثلت في مظهرين :

- أولهما تنافر صوتي الفعل.
- وثانيهما اتجاه صوتي الفعل إلى التصرف على باب صرفي بعينه

رأيت تمحيص هذه الارتباطات ، وتسجيل ما يمكن أن يمثل قواعد عامة تحكم ارتباط صوتي المضعف ببابه الصرفي، أو تنافرهما، وهي القواعد الصرفوصوتية للفعل الثلاثي المضعف morpho-phonemic rules.

أهداف البحث

1- رصد القواعد التي تحكم تنافر صوتي الفعل الثلاثي المضعف.

2- تصنيف هذه القواعد، وتحديد مدى شمولها أو اقتصارها على أصوات وأحياز بعينها.

3- تحديد القواعد العامة الشاملة لتنافر صوتي هذا النوع من الأفعال، وكذلك القواعد المختصة بأصوات ومخارج دون غيرها.

مادة الدراسة

اعتمدت الدراسة القاموس المحيط للفيروزابادي لغزارة مادته مع اختصاره، وحرصه على ضبط حروف كلماته بالشكل، إلى جانب التزامه بتحديد الباب الصرفي لأفعاله، بربطها بأوزان الأفعال المعروفة.

واستقصت الدراسة الأفعال الثلاثية الصحيحة المضعفة العين واللام به ، واتخذتها كلها عينة للبحث.

خطوات البحث

- ارتكزت الدراسة على جدول شامل يستقصي الأفعال الثلاثية المضعفة بالقاموس المحيط ، ويحدد أبوابها الصرفية: الجدول رقم (1).
- واستخرجت منه جدولا آخر يقتصر على تحديد الصوتين المتنافرين، ويختص بتحديد أصوات فاء المضعف التي تتنافر مع عينه ولامه : الجدول رقم (2).
- ثم استخرجت جدولا ثالثا يحدد أصوات عين المضعف ولامه وأثرهما في تنافر صوتيه : الجدول رقم (3).

المصطلحات:

- **المخرج: Point of articulation**

  هوالنقطة التي يلتقي عندها عضوان أو أكثر من أعضاء النطق ليمر هواء الزفير بينهما، ويتشكل الصوت.

- **الحيز: Range of articulation**

  مساحة تشتمل على أكثر من مخرج، وتكون المخارج فيها متقاربة.

- **الصوت المجهور، والصوت المهموس. Voiced & voiceless**

- **الإطباق، والانفتاح. Velarization & Non velarization**

- **الأصوات المتوسطة (الموائع): Liquids**

  تنطق بالتقاء تام لعضوين من أعضاء النطق، ولكن النفس يجد مسربا إلى الخارج، فيمر الهواء دون أن يحدث صفيرا أو حفيفا مسموعا.

- **اتبع البحث ترتيب الخليل للأصوات الصامتة، مع الأخذ برأي سيبويه في تقسيم الأصوات الحلقية، فأضاف إليها الهمزة.**

تقسيم الصوامت في البحث:

- أصوات الحلق: (أ - هـ - ع - ح - غ - خ) [Pharyngeal].
- صوتا اللَّهَاةِ وَالْحَنَكِ الأَعْلَى : (ق [uvular] - ك [velar]).
- الأصوات الشجرية:(ج - ش - ض) [postalveolar]

[الشجر: جوف الفم بين سقف الحنك واللسان]

- الأصوات الأسلية:(ص - س - ز)[alveolar] وتسمى أصوات الصفير sibilants

[تبدأ من أسلة اللسان: وهي مستدق طرفه]

- الأصوات النطعية: (ط - ت - د)[dental] [تبدأ من نطع (ظهر) الغار الأعلى]
- الأصوات اللثوية: (ظ - ث - ذ) [بين الأسنانية interdentals]
- الأصوات الذلقية:(ر - ل - ن) [وهي الأصوات المتوسطة أو **الموائع** liquids]

[تبدأ من ذلق اللسان: وهو تحديد طرفي حد اللسان]

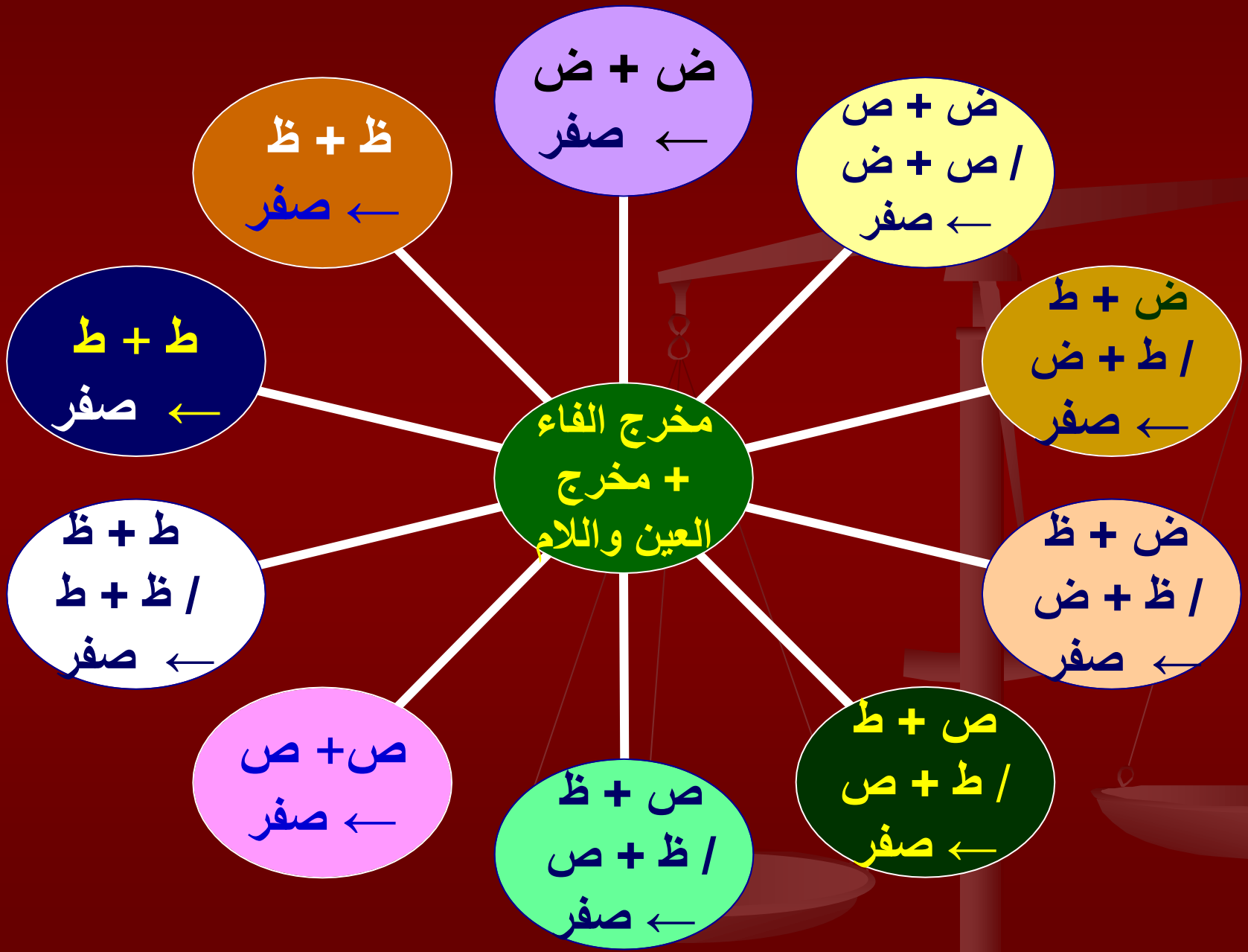
- الأصوات الشفهية : (ف - ب - م) [labial]

نتائج البحث
قواعد تنافر صوتي الفعل الثلاثي المضعف

أولا : القواعد العامة:

- لا يقع (الهمزة) أقصى الحلقي المجهور عينا ولاما للفعل الثلاثي المضعف.
- يتنافر صوتا المضعف إذا اتفقا في صفة الإطباق.
- تتنافر أصوات الحيز الواحد: فلا يقع أحدها فاءً والآخر عينا ولاما للمضعف.

يتنافر صوتا المضعف إذا اتفقا في صفة الإطباق



أحوال تصرف الأصوات النطعية والشفهية

فاء المضعف	عينه ولامه : نطعي			عينه ولامه : شفهي		
	ط	ت	د	ف	ب	م
ط	-	-	-			
ت	-	-	-	-		
د	-	-	-			
ف	-			-	-	-
ب				-	-	-
م				-	-	-

تتنافر أصوات الحيز الواحد: فلا يقع أحدها فاءً والآخر عينا ولاما للمضعف.

تابع: قواعد تنافر صوتي الفعل الثلاثي المضعف

قواعد خاصة بأصوات بعينها:

أولا : تنافر بتأثير المخارج أو الأحياز:

- تتنافر الأصوات الأسلية (ص- س- ز) مع اللثوية (ظ – ث - ذ) أيا كان موقعها من الفعل. (تنافر أحياز)

- يتنافر صوتا أقصى الحلق (أ – هـ) – عينا ولاما - مع أصوات الأحياز الوسطية . (تنافر مخرج مع أحياز)
- يتنافر (التاء) النطعي – فاء – مع الأحياز الوسطية.
- يتنافر (الظاء) اللثوي – فاء – مع الأحياز الوسطية.

(تنافر مخرج مع أحياز)

أحوال تصرف الأصوات الأسلية واللثوية

فاء المضعف	عينه ولامه : أسلي			عينه ولامه : لثوي		
	ص	س	ز	ظ	ث	ذ
ص	_	_	_	-	-	-
س	_	_	_	-	-	-
ز	_	_	*	-	-	-
ظ	_	_	_	-	-	-
ث	_	_	_	-	-	-
ذ	_	_	_	-	-	-

- تتنافر أصوات الحيز الواحد: فلا يقع أحدها فاءً والآخر عينا ولاما للمضعف.
- تتنافر الأصوات الأسلية واللثوية أيا كان موقع الصوت من المضعف.

تابع: القواعد الخاصة بتنافر أصوات بعينها في الفعل المضعف

ثانيا : تنافر بسبب ارتباط الحيز أو المخرج مع الصفة :

- يتنافر (ق) اللهوي - فاء - مع (غ - خ) أدنى الحلقيين.

(اتفاق في الاستعلاء)

- تتنافر الأصوات النطعية (ط – ت - د)- فاء- مع (الزاي) الأسلي و(الذال) اللثوي.

(تنافر أصوات حيز مع المجهور المنفتح من حيز مجاور)

- تتنافر الأسليات (ص- س- ز) – فاء - مع (ش– ض) الشجريين المستطيلين. (تنافر أصوات حيز مع صوتي حيز مجاور لهما صفة خاصة)

- تتنافر الأصوات الذلقية (ر- ل - ن) – فاء - مع الحلقيات المجهورة (أ- ع - غ).

(اتفاق في الجهر)

تصرف الأصوات الأسلية مع الأصوات الشجرية

فاء المضعف	عين المضعف ولامه : شجري		
	ج	ش	ض
ص	صج	—	—
س	سج	—	—
ز	زج	—	—

- تتنافر الأصوات الأسلية – فاء - مع الشجريين المستطيلين (من حيز مجاور، ولهما صفة خاصة).

تصرف الأصوات الذلقية ـ فاءً ـ مع الأصوات الحلقية

فاء الفعل	أصوات الحلق عينا ولاما للمضعف					
ذلقي مجهور	أ مجهور	هـ مهموس	ع مجهور	ح مهموس	غ مجهور	خ مهموس
ر مكرر	صفر	صفر	صفر	صفر	صفر	
ل جانبي	صفر		صفر		صفر	
ن خيشومي	صفر	صفر	صفر		صفر	

- تتنافر الذلقيات – فاءً ـ مع الحلقيات المجهورة.

تابع: القواعد الخاصة بتنافر أصوات بعينها في الفعل المضعف

ثالثا : تنافر بتأثير المخرج مع صفات الصوتين :

عند تطابق صفات الاحتكاك والهمس والانفتاح والاستفال:

- تنافر (الثاء) اللثوي – فاء – مع (الفاء) الشفهي.
- تنافر (الفاء) الشفهي – فاء – مع (السين) الأسلي.
- تنافر (الهاء) الحلقي – فاء - مع (الثاء) اللثوي.
- تنافر (الشين) الشجري – فاء – مع (الثاء) اللثوي.

أثر تطابق صفات الاحتكاك والهمس والانفتاح والاستفال في التنافر

حيز الصوت	فاء الفعل	عينه ولامه: أسلي			عينه ولامه: لثوي			عينه ولامه: شفهي		
		ص	س	ز	ظ	ث	ذ	ف	ب	م
أسلي	ص	–	–	–	–	–	–			
	س	–	–	–	–	–	–			
	ز	–	–	*	–	–	–			
لثوي	ظ	–	–	–	–	–	–		–	–
	ث	–	–	–	–	–	–	–		
	ذ	–	–	–	–	–	–			
شفهي	ف		–					–	–	–
	ب							–	–	–
	م						–	–	–	–

أثر تطابق صفات الاحتكاك والهمس والانفتاح والاستفال في التنافر

حيز الصوت	فاء الفعل	عينه ولامه: شجري			عينه ولامه: لثوي		
		ج	ش	ض	ظ	ث	ذ
أقصى الحلق	أ				–		–
	هـ				–	–	
شجري	ج	–					
	ش		–	–		–	
	ض		–	–	–	–	–

- يتنافر (الهاء) الحلقي – فاء - مع (الثاء) اللثوي. (تطابق الصفات)
- يتنافر (الشين) الشجري – فاء – مع (الثاء) اللثوي. (تطابق الصفات)

خاتمة

بهذا يكون البحث قد حقق أهدافه بعد أن:

1- رصد القواعد التي تحكم تنافر صوتي الفعل الثلاثي المضعف.

2- صنف هذه القواعد، محددا مدى شمولها، أو اقتصارها على أصوات وأحياز بعينها.

3- حدد القواعد العامة الشاملة لتنافر صوتي هذا النوع من الأفعال، والقواعد المختصة بأصوات ومخارج دون غيرها.

الجدوى التطبيقية لهذه الدراسة

- في اللسانيات التطبيقية :
- الصناعة المعجمية : Lexicography

توضيح القواعد التي يسلكها المعجم العربي في تأليف أصوات وحداته وتنافرها، وتعليل ذلك.

- تعليم اللغة : Language learning

معرفة الأصوات والأحياز التي لا تجتمع معا تسهم في تعليم اللغة لغير الناطقين بالعربية.

- المصطلحية : Terminology

توضيح قواعد التآلف والتنافر في تكوين الكلمة العربية يفيد في وضع ركائز لسك المصطلحات الجديدة، وتعريب المصطلح الأجنبي.

- في العمل المعجمي الحاسوبي :
- بناء قاعدة بيانات معجمية :

Lexical database construction

باستقصاء الكلمات والصيغ الممكنة وتلك الممتنعة، وإحصاء ذلك آليا؛ للتوصل إلى القواعد الصوتية ، وكذلك الصرفية الصوتية التي تحكم المعجم.

- تعرّف الكلام: Speech recognition

ببناء نماذج تشمل قواعد التتابعات الممكنة صوتيا، والتتابعات غير الممكنة؛ مما يسهل عملية الإدراك الآلي للأصوات.

شكرا لحسن استماعكم

جدول رقم (3)

الأفعال المضعفة الثلاثية التي لا تقع في العربية بسبب تنافر أصواتها

عين الفعل ولامه	حيز أصوات فاء المضعف																					
	الحلق						اللهاة		الأحياز الوسطية										الذلقية			الشفهية
	أقصاه		وسطه		أدناه		لهاة	حنك	شجر القم			الأسلة			تطع الغار			اللثة			حروف الذلاقة			الشفتان		
	أ	هـ	ع	ح	غ	خ	ق	ك	ج	ش	ض	ص	س	ز	ط	ت	د	ظ	ث	ذ	ر	ل	ن	ف	ب	م
همزة	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
هاء	*	*	–	–	–	–			–	–	*	–	–	–	–	–	–	–	–	–	–	–		–			
عين	–	*	–	–	–	**						–	–	–				–		–			–	–			
حاء	*	–	–	–	–	–		–			–					–		–	–		–						
غين	–	–	–	–	–	–	–	–	–		–		–	–	–	–	–	–	–		–	–	–	–			–
خاء	–	–	–	–	–	–	–	–							–		–	–	–	–							–
قاف	–						–	–							–	–		–	–	–							
كاف					–	–	–	–	–						–			–		–			–				

جدول رقم (2)

الأفعال المضعفة الثلاثية التي لا تقع في العربية بسبب تنافر أصواتها (مظللة)

فاء الفعل	حيز أصوات عين الفعل ولامه																					
	الحلق						اللهاة		شجر الفم			الأسلة		تطع الغار			اللثة			حروف الذلاقة			الشفتان		
	أ	هـ	ع	ح	غ	خ	ق	ك	ج	ش	ض	ص	س	ط	د	ت	ظ	ث	ذ	ر	ل	ن	ف	ب	م
همزة	-	(1)	-	(2)	-	-	-										-		-							
هاء	-	(3)	(4)	-	-	-								-			-	-								
عين	-	-	-	-	-	-													-							
حاء	-	-	-	-	-	-																				
غين	-	-	-	-	-	-		-	-								-						-			
خاء	-	-	(5)	-	-	-		-			-							-								
قاف	-				-	-	-	-	-								-									
كاف	-			-	-		-	-			-			-												

(1) تص القاموس المحيط ولسان العرب، مادة (أ هـ هـ) على أن : " الأهة: التحزن، وقد أة أَهًّا وأَهَّة" ويلحظ أنه حكاية صوت.
(2) أورد القاموس الفعل (أحّ) بمعنى: سعل. وتص اللسان على أنه حكاية صوت، مادة (أ ح ح):" (أحّ) حكاية تتحنح أو توجع".
(3) الفعل (هه) حكاية صوت.
(4) تص القاموس على أن الفعل (هغ) لغة في (هاع).
(5) أورد القاموس الفعل (خغ). وجاء باللسان ما يشير إلى أنه حكاية صوت القهد إذا اتبهر، ويشكك في صحته.

تابع : جدول رقم (2)

الأفعال المضعفة الثلاثية التى لا تقع فى العربية بسبب تنافر أصواتها (مظللة)

فاء الفعل	حيز أصوات عين الفعل ولامه																							
	الحلق						اللهاة		شجر الفم			الأسلة			نطع الغار			اللثة			حروف الأذلاقة			الشفتان		
	أ	هـ	ع	ح	غ	خ	ق	ك	ج	ش	ض	ص	س	ز	ط	ت	د	ظ	ث	ذ	ر	ل	ن	ف	ب	م
جيم	-	-			-			-	-						-											
شين	-	-								-	-								-							
ضاد	-	*		-	-					-	-	-	-		-	-		-	-	-						
صاد	-	-	-						(1)	-	-	-	-	-	-			-	-	-			-			
سين	-	-	-		-					-	-	-	-	-	-	-		-	-	-						
زاى	-	-	-		-					-	-	-	-	**			-	-	-	-						
طاء	-	-			-		-	-	-		-	-		-	-	-	-	-	-	-						
ناء	-	-		-	-			-	-	-	-	-	-	-	-	-	-	-	-	-			-	-		
دال	-	-			-	-								-	-	-	-			-						

* أورد القاموس الفعل (ضهّ) وذكر أنه "لغة في ضاهاه". ولم يرد الفعل باللسان.

(1) القاموس: "صجّ: ضرب حديداعلى حديد فصوّتا، والصَّجُج بضمتين ذلك الصوت". اللسان: "صجج: أهملها الليث، وروى أبو العباس عن ابن الأعرابي: صجّ إذا ضرب حديدا على حديد فصوّتا. والصجيج ضرب الحديد بعضه على بعض". كما جاء باللسان: " الصخّ: الضرب بالحديد على الحديد".

** جاء بالتاج: " ززّ: أهمله جمهور المصنفين في اللغة، وإنما أورده بعض أئمة الصرف فيما استوت مادته في البناء كببة وشبهه، ومن ثم يبدو أن هذا الفعل من الأفعال المهملة أو المصنوعة لغرض تعليمي.

تابع : جدول رقم (2)

الأفعال المضعفة الثلاثية التي لا تقع في العربية بسبب تنافر أصواتها (مثلثة)

فاء الفعل	حيز أصوات عين الفعل ولامه																							
	الحلق						اللهاة		شجر الفم			الأسلة			نطع الغار			اللثة			حروف الذلاقة			الشفتان		
	أ	هـ	ع	ح	غ	خ	ق	ك	ج	ش	ض	ص	ز	س	ط	ت	د	ظ	ث	ذ	ر	ل	ن	ف	ب	م
ظاء	-	-	-	-	-	-	-	-	(1)	-	-	-	-	-	-	-	-	-	-	-					-	-
ثاء	-	-		-	-	-	-			(2)	-	-	-	-		-	-	-	-	-				-		
ذال	-	-	-			-	-	-		(3)	-	-	-	-	-	-	-	-	-	-						
راء	-	-	-	-	-				-						-			-			-	-				
لام	-		-		-					-	-										-	-	-			
نون	-	-	-		-			-										-			-	-	-			
فاء	-		-											-	-									-	-	-
باء	-																							-	-	-
ميم	-		-		-	-														-				-	-	-

(1) جاء بالتاج (ظ ج ج): " ظج: صاح في الحرب صياح المستغيث، قاله ابن الأعرابي.وقال أبو منصور: الأصل فيه (ضج)بالضاد، ثم جعل ضج في غير الحرب، و(ظج) بالظاء في الحرب

(2) جاء بالتاج (ث ش ش):" ثش:أهمله الجوهري وصاحب اللسان، وقال أبو عمرو: ثش سقاءه وفشه‌أخرج منه الريح، هكذا نقله عنه الصاغاني، وكأن الثاء بدل من الفاء"

(3) جاء بالتاج (ذ ش ش): " ذش الرجل،أهمله الجوهري والجماعة، ونقل الصاغاني عن ابن الأعرابي: أي سار، لغة في (دشّ) بالدال"

جدول رقم (3)

الأفعال المضعفة الثلاثية التي لا تقع في العربية بسبب تنافر أصواتها

عين الفعل ولامه	حيز أصوات فاء المضعف																					
	الحلق						اللهاة		الأحياز الوسطية										الذلقية			الشفهية
	أقصاه		وسطه		أدناه		لهاة	حنك	شجر الفم			الأسلة			نطع الغار			اللثة			حروف الذلاقة			الشفتان		
	أ	هـ	ع	ح	غ	خ	ق	ك	ج	ش	ض	ص	س	ز	ط	ت	د	ظ	ث	ذ	ر	ل	ن	ف	ب	م
همزة	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
هاء	*	*	–	–	–	–			–	–	*	–	–	–	–	–	–	–	–	–	–	–		–			
عين	–	*	–	–	–	**						–	–	–				–		–		–	–		–		
حاء	*	–	–	–	–	–		–			–					–		–	–		–						
غين	–	–	–	–	–	–	–	–	–		–		–	–	–	–	–	–		–	–	–	–	–			–
خاء	–	–	–	–	–	–	–										–	–	–	–	–						–
قاف	–						–	–							–	–		–		–	–						
كاف					–	–	–	–	–						–			–		–			–				

تابع جدول رقم (3)

الأفعال المضعفة الثلاثية التي لا تقع في العربية بسبب تنافر أصواتها

عين الفعل ولامه	حيز أصوات فاء المضعف																							
	الحلق						اللهاة		الأحياز الوسطية												الذلقية			الشفهية
	أقصاه		وسطه		أدناه		لهاة	حنك	شجر الفم			الأسلة			نطع الغار			اللثة			حروف الذلاقة			الشفتان		
	أ	هـ	ح	ع	غ	خ	ق	ك	ج	ش	ض	ص	س	ز	ط	ت	د	ظ	ث	ذ	ر	ل	ن	ف	ب	م
جيم					–		–		–			(1)			–	–		(2)								
شين										–	–	–	–	–		–		–	(3)			–				
ضاد						–		–		–	–	–	–	–	–	–		–	–	–		–				
صاد											–	–	–	–	–	–		–	–	–						
سين											–	–	–	–		–		–	–	–				–		
زاي										*		–	–		–	–	–	–	–	–						
طاء		–						–	–		–	–	–			–	–	–		–	–			–		
تاء											–		–			–	–	–	–	–						
دال																–	–	–	–	–						

تابع جدول رقم (3)

الأفعال المضعفة الثلاثية التي لا تقع في العربية بسبب تنافر أصواتها

عين الفعل ولامه	حيز أصوات فاء المضعف																							
	الحلق						اللهاة		الأحياز الوسطية										الذلقية			الشفهية		
	أقصاه		وسطه		أدناه		لهاة	حنك	شجر الفم			الأسلة			نطع الغار			اللثة			حروف الذلاقة			الشفتان		
	أ	هـ	ع	ح	غ	خ	ق	ك	ج	ش	ض	ص	س	ز	ط	ت	د	ظ	ث	ذ	ر	ل	ن	ف	ب	م
ظاء	–	–			–		–				–	–	–	–	–	–		–	–	–	–		–			
ثاء		–				–				–	–	–	–	–	–	–		–	–	–						
ذال	–		–								–	–	–	–	–	–	–	–	–	–				–		
راء																					–	–	–			
لام																					–	–	–			
نون												–				–			–			–	–			
فاء					*											–			–					–	–	–
باء																		–						–	–	–
ميم			–															–						–	–	–

Bel-Arabi Advanced Arabic Dependency Structure Extractor

- Michael Nashaat Nawar
- Mahmoud Nabil Mahmoud

10/7/2024

Agenda

- Problem Definition
- Related Work
- System Architecture
- System Limitations
- System Evaluation
- Demo

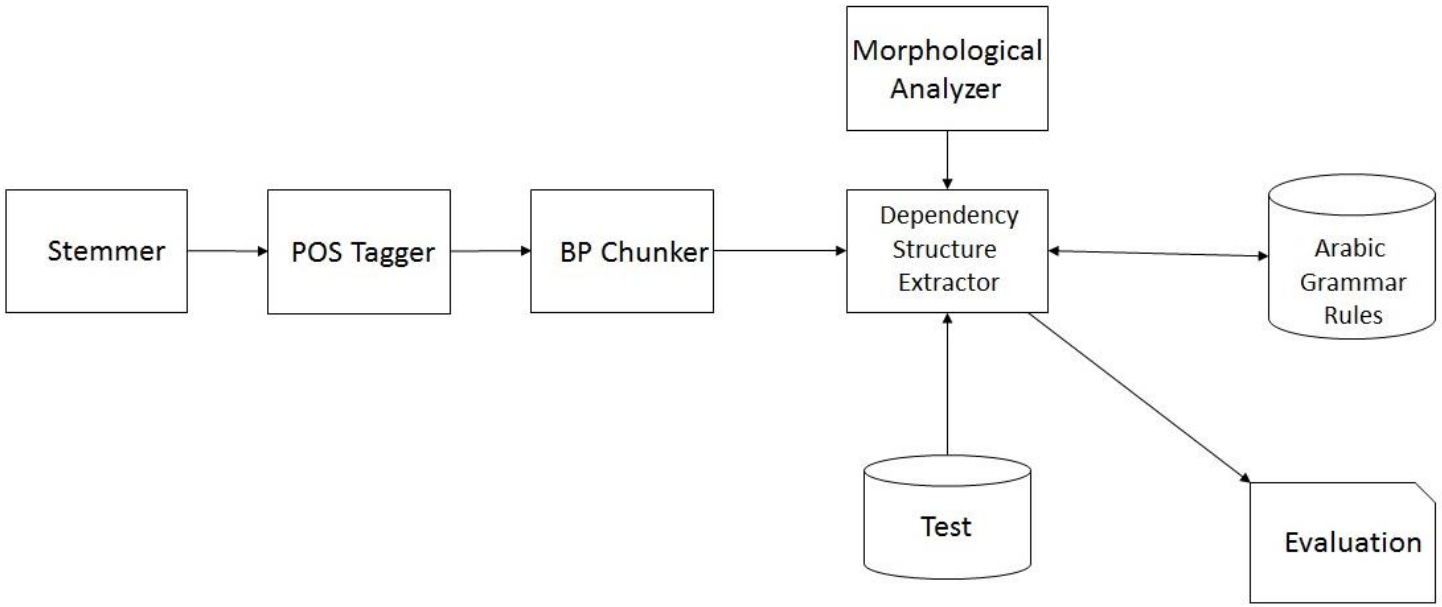
Problem Definition

- Limited work has been practiced on Arabic NLP.
- Dependency Structure Extraction is a complex task.
- Arabic dependency structure extractor can solve many problem such as automatic diacritics, Arabic sentences correction and accurate translation.

Related Work

- Al Daoud et al. propose a framework to automate the relation extraction of Arabic language verbal sentences.
- Attia built an Arabic parser using Xerox linguistics environment.
- Habash et al. construct The Columbia Arabic Treebank (CATiB).

System Architecture



System Limitations

- The system is assuming that sentence has been written correctly.
- The system assumes the verb as it is in the active voice.
- The dependency structure extractor does not prevent errors that are related to incorrect use of semantic meaning, means that the semantic analysis is not verified

System Evaluation Results

- We have generated 600 sentences consisting of 3452 tokens.

	Tags	Parses	Signs
Precision	0.9567	0.9575	0.9801
Recall	0.9422	0.9518	0.6426
F-measure	0.9504	0.9546	0.7230
Item Accuracy	0.9333	0.9409	0.9449

Tools

- Microsoft Visual Studio 2010

- QT Creator for Graphical user interface

Future Work

- Increasing the coverage of the morphological analyzer by using other data sources like Wikipedia Arabic dump.
- Using more corpora to train Stemmer, POS tagger, and Base Phrase Chunker.
- Increasing the coverage and the accuracy of the dependency structure extractor by writing more rules.

Demo

THANK YOU

Sentiment Analysis Improvement Using the Transformation of Colloquial Text to Standard Arabic

Fatma El-zahraa El-taher
Alaa El-Dine Ali Hamouda
Salah Abdel-Mageid

Agenda

- Introduction
- Problem Definition
- Proposed Solution
- System Evaluation
- Conclusion

Introduction

- Sentiment Analysis becomes very important due to the increase of on-line social-oriented content (e.g., user reviews, blogs, Facebook comments, tweets, etc)
- Although there is a lot of work in sentiment analysis in different languages, there is a limited research in sentiment analysis for Arabic content.

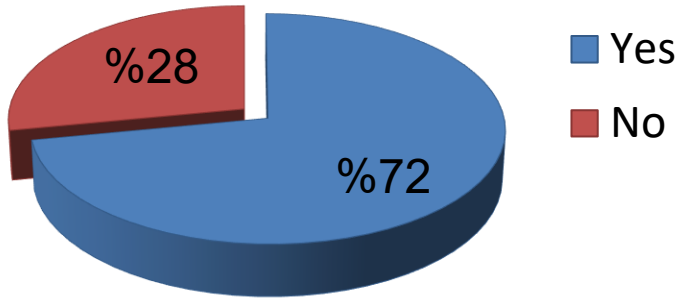
Introduction

- Users have become more interested in following news and governmental pages on Facebook

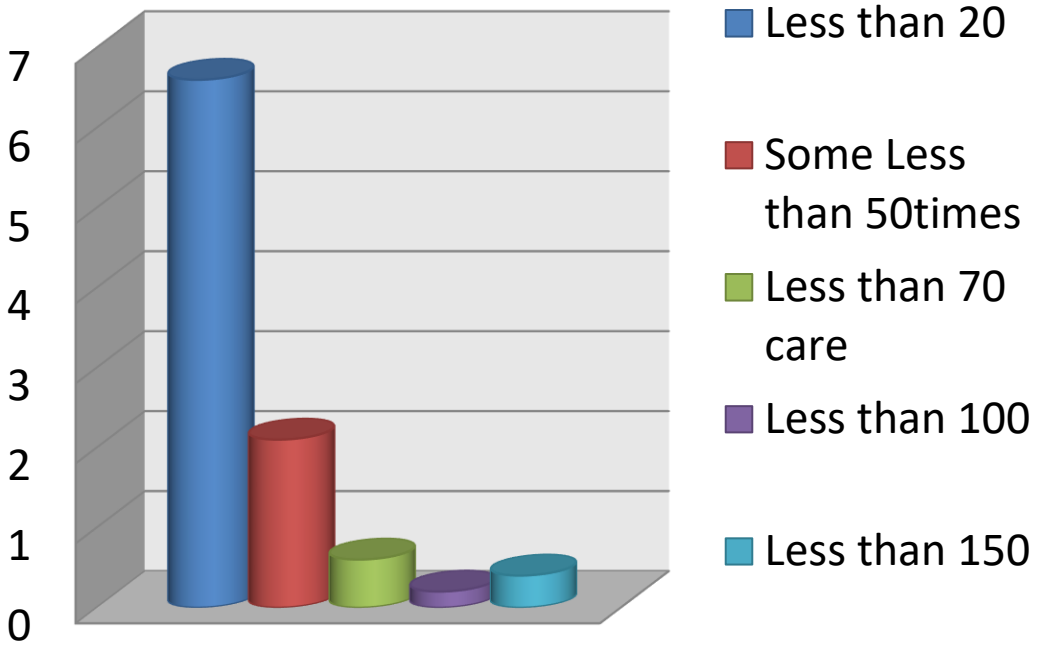
رصد	4,197,917 likes
الصفحه الرسميه لرئاسه مجلس الوزراء	1,294,174 likes
شبكه اخبار مصر	1,174,196 likes

Survey

Do you Follow the popular pages?



How many comments do you usually read?



***We made a survey with a population 497 of Facebook users**

Agenda

- Introduction
- Problem Definition
- Proposed Solution
- System Evaluation
- Conclusion

Problem Definition

الصفحة الرسمية لرئاسة مجلس الوزراء المصرى

1,298,438 like this

November 26 at 7:56pm near Cairo

اجتمع الدكتور حازم الببلاوى رئيس مجلس الوزراء اليوم بعدد من ممثلي جبهة الإنقاذ وممثلي الشباب بحضور وزير التضامن الاجتماعي الدكتور أحمد البرعي، وذلك لمناقشة تطورات الأوضاع السياسية والاقتصادية، وقد تناولت المناقشات أيضاً موضوع قانون تنظيم الحق في التظاهر والأحداث التي وقعت ظهر اليوم، وطالب المجتمعون بالإفراج عمن تم احتجازهم اليوم أثناء مشاركتهم في التظاهرات احتجاجاً على القانون، وقد وعد رئيس الوزراء بمتابعة ما تسفر عنه تحقيقات النيابة العامة في هذا الشأن توصلاً للاستجابة لهذا المطلب.

وقد أبدى الحاضرون عدداً من الاعتراضات والتحفظات على بعض مواد القانون، وتم الاتفاق على تشكيل لجنة مجتمعية مشتركة لدراسة هذه الأراء.

Like · Comment · Share 69

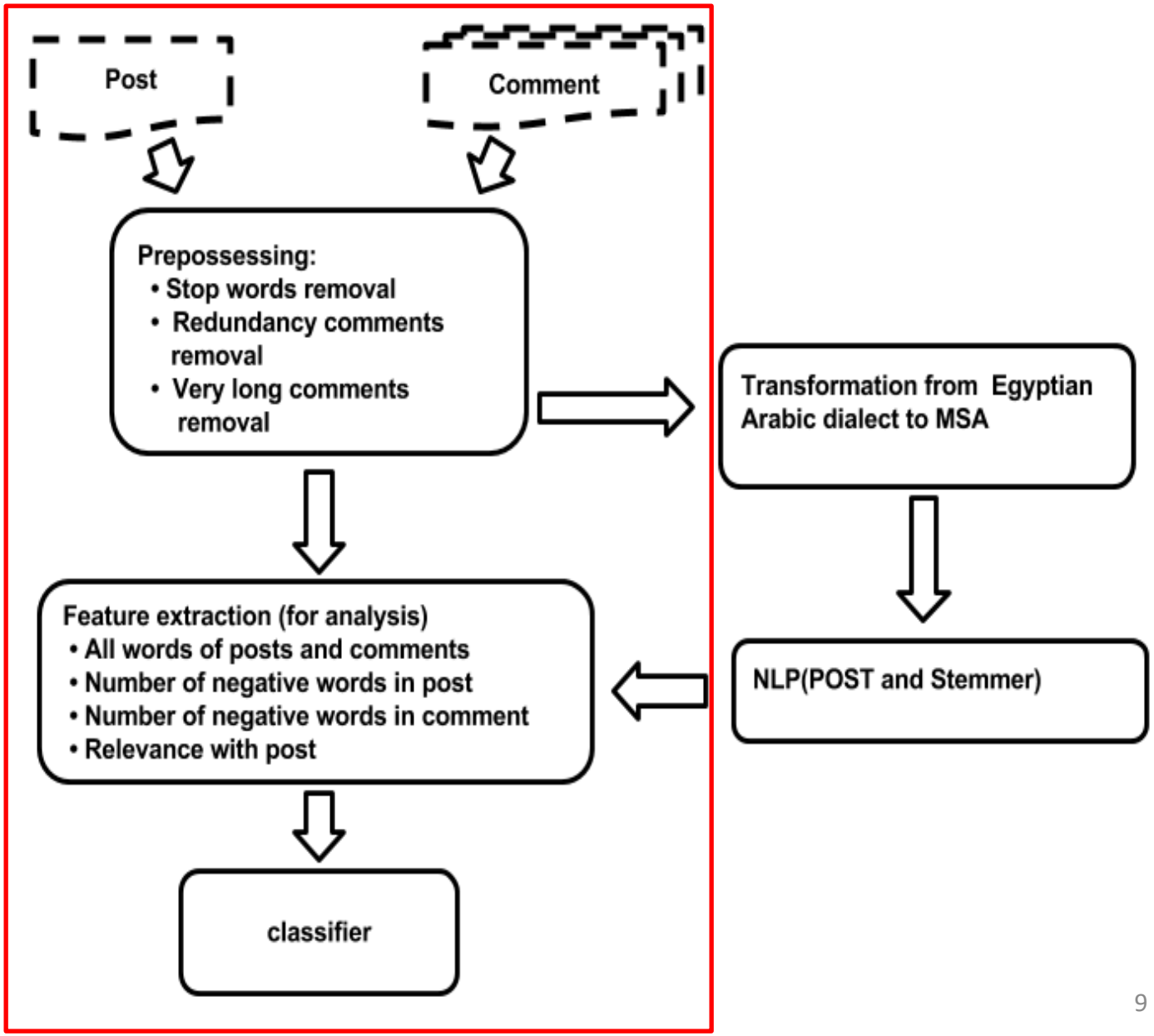
75 people like this.

322 comments

Agenda

- Introduction
- Problem Definition
- Proposed Solution
- System Evaluation
- Conclusion

Sentiment Analyzer Block Diagram



Data Collection

Prepare collection of comments as corpus

- For training data set, we collected comments from **news and governmental Facebook pages**
- Corpus size is 1200 comments collected from 49 posts.
- Comments are divided into **3 groups**; supportive comments, attacking comments and neutral comments.

Proposed Solution (Cont’)

Preprocessing Stage:

1. Stop words removal like (اللى ,دى ،ده)
2. Special character and redundancy letters removal like (@,!, % ,منقووووول)
3. Long comments removal (ignore comments with number of words more than 150 words)

Features Extraction

1. All Words in Posts and Comments Feature

Example :

Post: الجنزوري يلتقي بالشيخ حسان لبحث الاستغناءعن المعونة

Comment1:ربنا يوفقك يا شيخ حسان

Comment2:أستغنواعن المعونة مع أنفسكم

	شيخ	أنفسكم	استغنوا	يوفقك	ربنا	المعونة	الاستغناء	حسان	الشيخ	يلتقي	الجنزوري
Comment1	N	C	C	N	N	M	M	H	M	M	M
Comment2	C	N	N	C	C	H	M	M	M	M	M

“**C**“ word is not in the post or the comment. “**M**” word is in the post only.

“**N**” word is in the comment only. “**H**” word is in both of the post and the comment.

Features Extraction (Cont')

2. Number of Negation Words in the post

It is a measure for the degree of negation in the post

$$\frac{\text{Number of negative words in the post}}{\text{length of the post}}$$

3. Number of Negation Words in the comment

It is a measure for the degree of negation in the comment

$$\frac{\text{Number of negative words in the comment}}{\text{length of the comment}}$$

Features Extraction and Classification

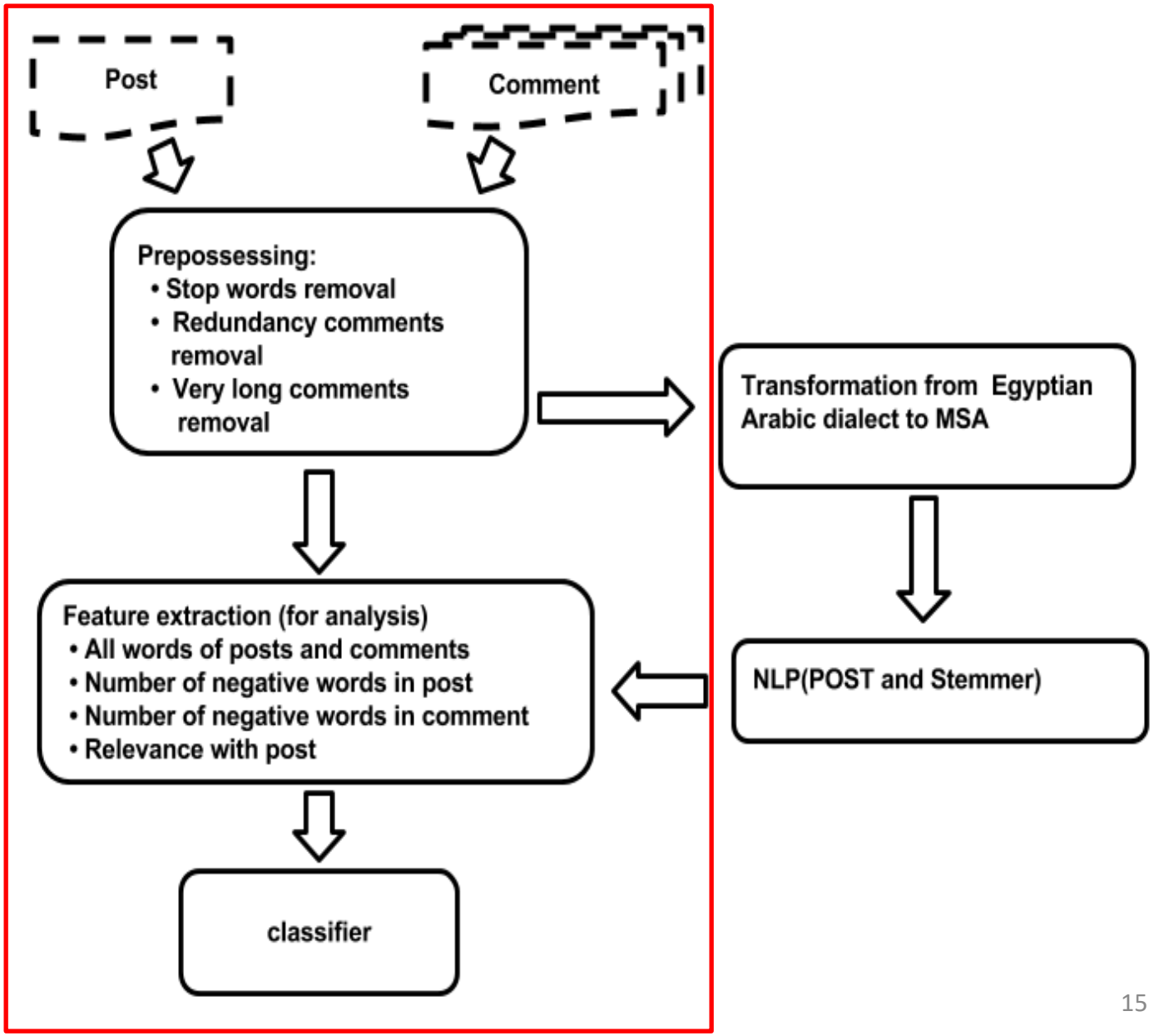
4. Relevance with post

$$Tf = F\ (1)$$

F is the number of occurrences of the word . Then the relevance is calculated using Cos function.

Then we apply SVM on these features.

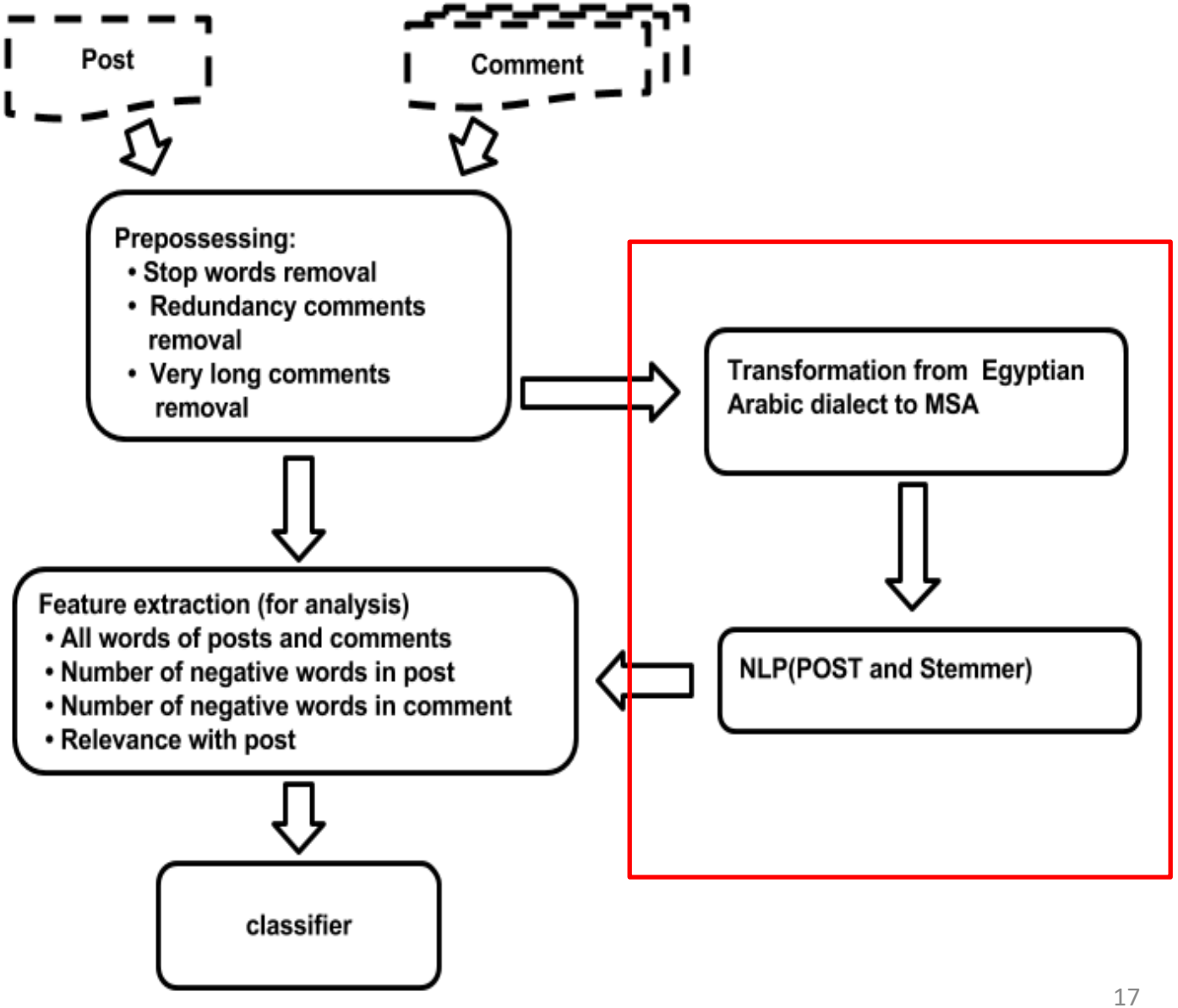
Sentiment Analyzer Block Diagram



System Evaluation

	Egyptian Comments	
	Precision	Recall
Attacking	59.8%	95.1%
Neutral	42.1%	4.1%
Supporting	55.7%	20.5%
Average	55.8%	59.1%
F-Measure	50.3%	

Sentiment Analyzer Block Diagram



Some Transformation Rules

- Remove the suffix (ش) from the end of negation verb.

  ما اعرفش ⟵ لا أعرف like

- Replace the letters (ح، هـ) from the verb with ((س, سوف

- Remove the prefix (أن, أت)from the passives verb

  انضرب ⟵ ضرب Like

New Features

1. Part of Speech Tagging

POS Tagging segments comment to words and gives each word a tag. In this case, a word with a tag is used as a feature. So we replace a word "يلتقى" with "يلتقي/VBP ".

2. Stem

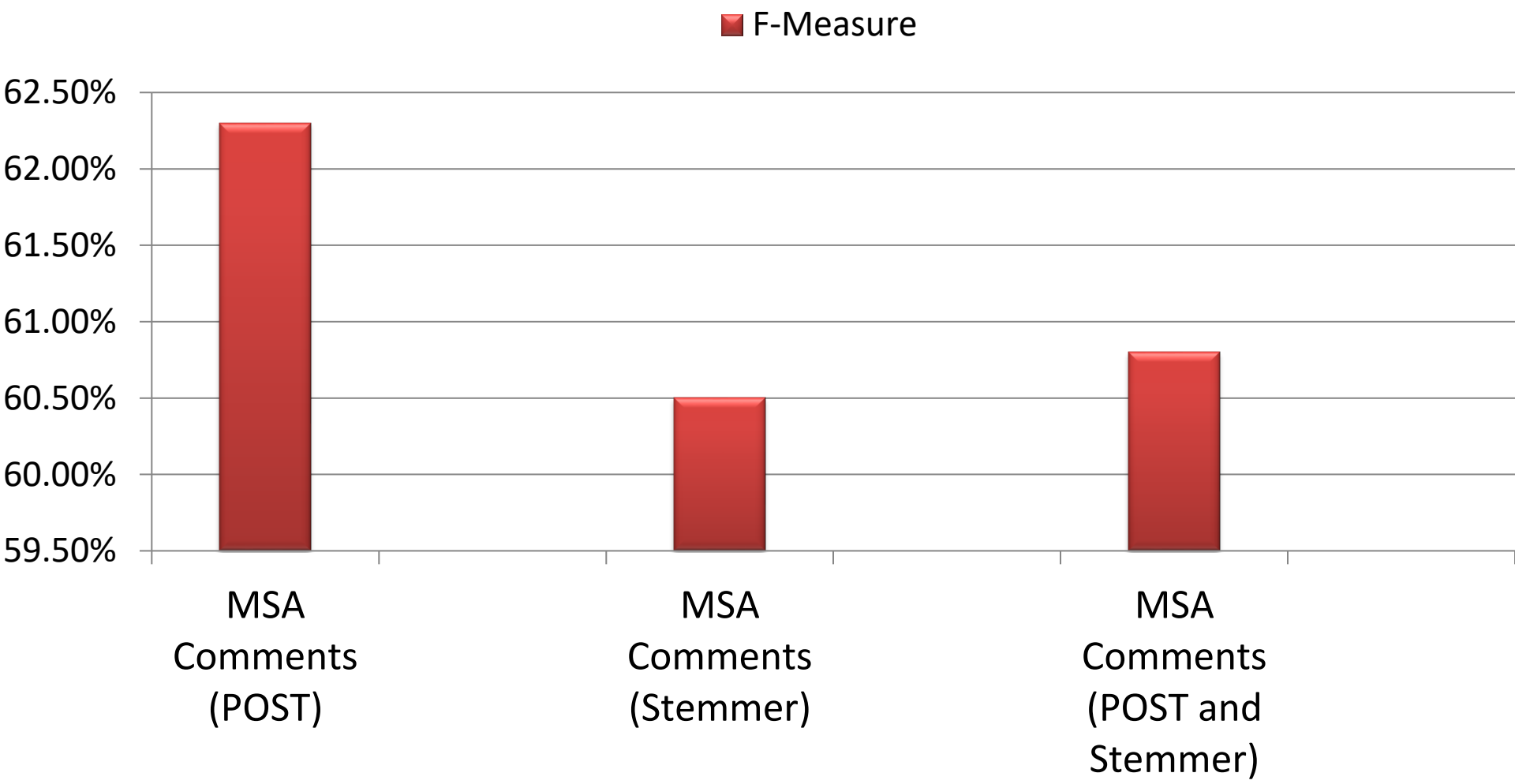
stemmer takes an Arabic word and returns the stem of it. In this case, the stem of the word is used as a feature. So we replace a word "يلتقى" with "لقي".

We used Stanford POST and Khoja's Stemmer

Agenda

- Introduction
- Problem Definition
- Proposed Solution
- System Evaluation
- Conclusion

System Evaluation



Agenda

- Introduction
- Problem Definition
- Proposed Solution
- System Evaluation
- Conclusion

Conclusion

- We construct corpus for supportive, attacking, and neutral comments with regard to different posts.
- Then we apply SVM classifier on Egyptian Arabic dialect and on the transformed comments into MSA after applying **POST** and **stemming**.
- The performance of the system improves by using the **POST and stemmer**.
- By applying the system in Egyptian comments , the performance of the system reaches 50.3%
- The best result is obtained by using **POST** on MSA Comments. We could reach up to **63.5%** of accuracy on the test set.

Thank you

هنسة عين شمس

جمعية هندسة اللغة

2013-12-11

فـروع الإنسانيـــات الجـدد وحوسبـــة اللغـــة

د. نبيل علي

INSPIRE OR EXPIRE

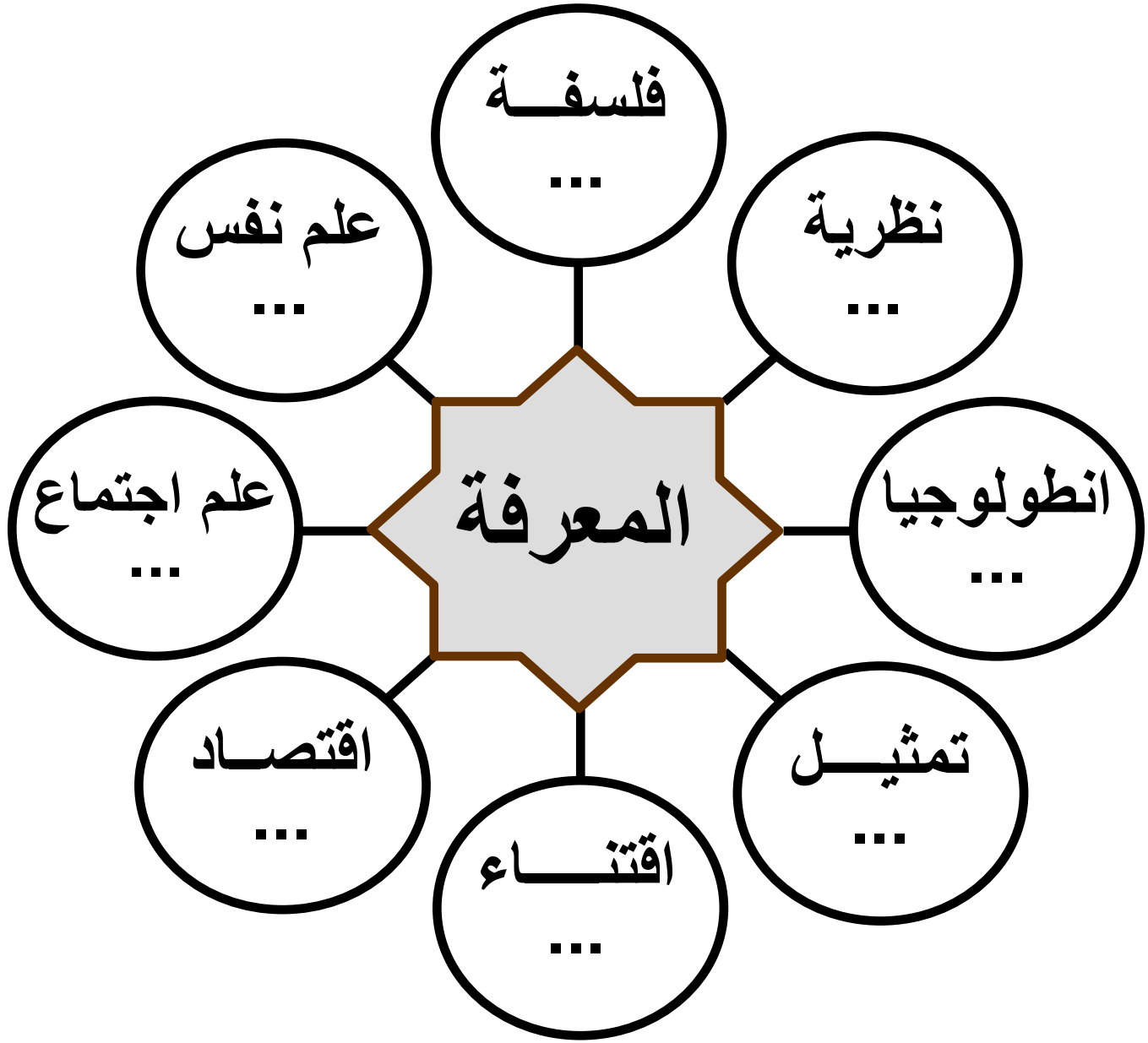
الإطار العام

1. الإنسانيات : النقلة النوعية
2. موسم الهجرة إلى الجمعي
3. حوار الشبكات
4. ثورة البيانات ومنهجيات حل المشكلات

علـــوم الانسانيـــات تجـــدد جلدهـــا ..!

علـــم الاجتمـــاع	المعرفي	SOCIOLOGY	COGNITIVE
علـــم النفـــــس	المعرفي	PSYCHOLOGY	COGNITIVE
علـــم اللغويـــات	المعرفي	LINGUISTICS	COGNITIVE
علـــم التاريـــخ	المعرفي	HISTORY	COGNITIVE
علـــم التربيـــة	المعرفي	PEDAGOGY	COGNITIVE
علـــم الجمـــال	المعرفي	AESTHETICS	COGNITIVE
علـــم الاقتصـــاد	المعرفي	ECONOMY	COGNITIVE

معـــــارف المعرفـــــــة



الذكاء الاصطناعي يلوذ بالفلسفة وعلوم والإنسانيات

- الذكاء الاصطناعي يستعصي على التطور من خلال تراكم التحسينات المتدرجة INCREMENTAL
- من خلال الفلسفة وعلوم الإنسانيات سنكتشف كم هي ضيقة نظرتنا للذكاء الاصطناعي
- لن يقلل ذلك من جاذبية أهل حوسبة اللغة في سوق العمل بل على العكس سوف يعززها
- حلم أصل الذكاء الاصطناعي هو محاكاة ما يجرى داخل المخ البشري، وإن تعذر ذلك حاليا فعلينا أن اقتفاء تجلياته المحسوسة ومظاهر سلوكه المختلفة

TIME GO COLLECTIVE موسم الهجرة إلى الجمعي

• COLLECTIVE INTELLIGENCE	• الذكاء الجمعي
• COLLECTIVE FILTERING	• الترشيح الجمعي
• COLLABORATIVE LEARNING	• التعليم التعاوني
• COLLABORATIVE KNOWLEDGE GENERATION	• توليد المعرفة تعاونيا
• COLLABORATIVE PROGRAMMING (OPEN SOURCE)	• تطوير البرامج تعاونيا
• CROWD SOURCING	• احتشاد المصادر
• SOCIAL SEARCH ENGINE	• محرك البحث الاجتماعي
• COLLABORATIVE CONSUMPTION	• الاستهلاك التعاوني
• PARTICIPATORY PLANNING	• التخطيط التشاركي

ما السر وراء كون الكثير أشد فطنة من القليل

WAY THE MANY IS SMARTER THAN THE FEW

حكمة الاحتشاد **CROWD WISDOM**

الاحتشاد من الغوغائية إلى الحكمة

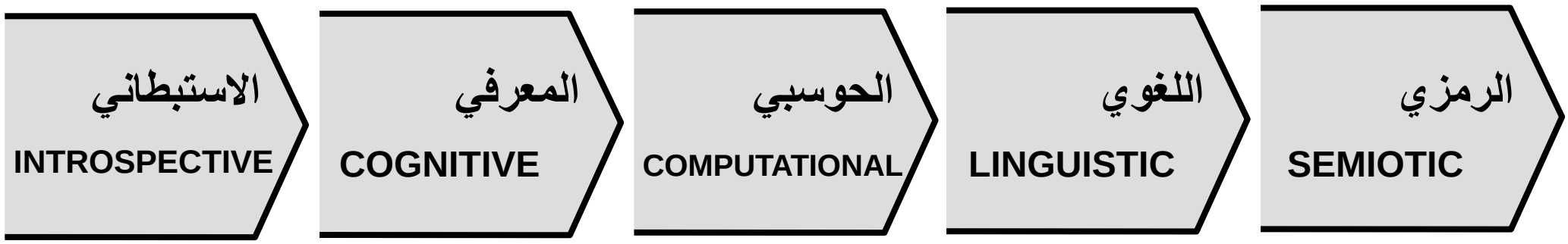
دعنا ننفض الضوضاء عن الظاهر المخادع الزائغ للكشف عن النظام الكامن في جوفه

التعلم العميق هو أداتنا للسيطرة على العشوائية السطحية للبيانات للكشف عما يعتملوا في جوفها من علاقات

ترديدات ثنائية الفردي والجمعي

النفسي PSYCHOLOGICAL	الاجتماعي SOCIOLOGICAL
استظهار EXTERNALIZE	استبطان INTERNALIZE
السردي NARRATIVE	الظاهراني PHENOMENOLOGICAL
الذاتي SUBJECTIVE	الموضوعي OBJECTIVE
الميكرو MICRO	الماكرو MACRO

سلسلة من النقلات النوعية



SEARCH ENGINES TO WHERE? محركات البحث إلى أين؟

مدخل البحث SEARCH ENTRY	لغويـــــا LINGUISTICS	حوسبة اللغـــــة LANGUAGE COMPUTATION
كلمات مفتاحية KEYWORDS	الصـــــرف MORPHOLOGY	معالجة الصرف آليا MORPHOLOGICAL PROCESSING
نصـــــي TEXTUAL	التركيـــب SYNTAX	الإعراب الآلـــــي AUTOMATIC PARSING
مفهومـــي CONCEPTUAL	الدلالـــــة SEMANTICS	الفهم الاتوماتي (ضحل/ عميق) AUT. UNDERSTANDING (SHALLOWLY/ IN-DEPTH)
اجتماعـــي SOCIAL	البرجماتيـــة PRAGMATICS	هندسة التخاطـــب CONVERSATIONAL ENGINEERING

حـــــوار الشبكات

NETWORK DIALOGUE

الشبكات الاجتماعيــة	الشبكات الأعصابية
SOCIAL NETWORK	NEURAL NETWORK

NETWORK DIALOGUE

حـــــوار الشبكات

الشبكات الأعصابية NEURAL NETWORK	الشبكات الدلاليـــــة SEMANTIC NET	الشبكات الاجتماعيــة SOCIAL NETWORK
انفجـــار البيانـــات DATA EXPLOSION	البنى النحوية المعجمية LEXICOS SYTACTICALS STRUCTURE	التفاعل الاجتماعي SOCIAL INTERACTION
التعلـــــم العميـــق DEEP LEARNING	السمـــات الدلاليـــة SEMANTIC FEATURES	علاقات التواصل البينية INTER- CONNECTIVITY

لا تفوق قدرة الإنسان على حل المشكلات إلا قدرته على خلق مشكلات جديدة

كورت جودل: مبدأ عدم الاكتمال

حل المشكلة معروف مسبقا

PROGRAM	=	ALGORITHM	+	DATA

حل المشكلة غير معروف

DEEP LEARNING	=	•NEURAL NET •DATA MINING	+	PREVIOUS DATA

التنقيب فـــي مناجـــم البيانـــات

DATA MINING

في عصرنا الرقمي افعالنا ورغباتنا وميولنا مسجلة إلكترونيا

- توقع الجرائم أوقاتها وأماكنها
- وقوع الحوادث
- تقلبات الأسواق
- معدلات الاستهلاك
- أنماط الطلب
- احتمالات الإصابة بالأمراض
- بيانات الفلك ومواقع النجوم

بيانـات ← **DM** → استخلصات

QUANTITATIVE METHODS

MATH. EQUATIONS

NEURAL NET

REGRESSION

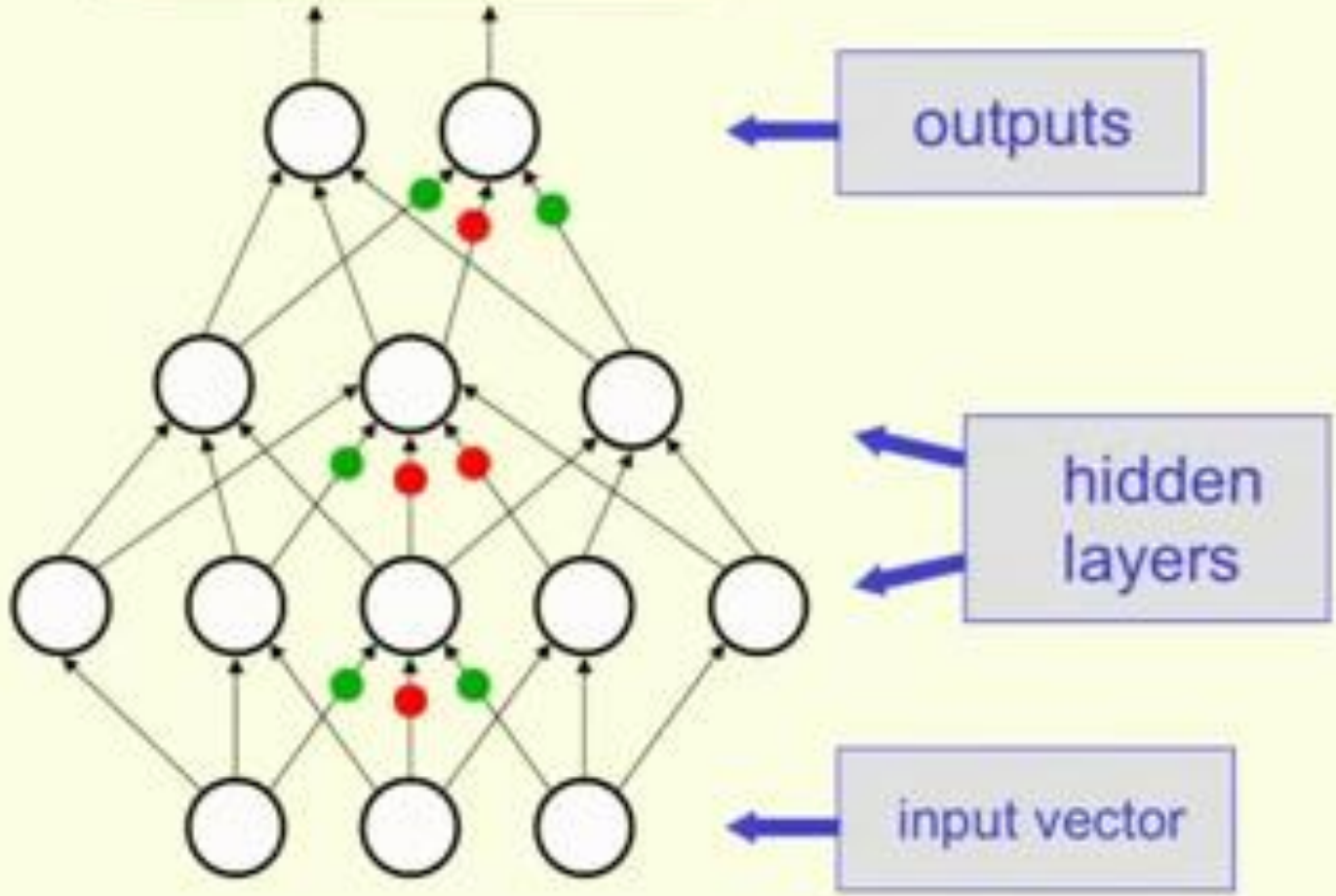
CLUSTERING

...

Deep neural networks (~1985)

Back-propagate error signal to get derivatives for learning

Compare outputs with correct answer to get error signal



طرق إيجاد حلول المشكلات

PROBLEM	SOLUTION	METHODOLOGY
KNOWN	KNOWN BY HUMANS	EXPERT SYSTEMS
KNOWN	AUTOMATIC	ALGORITHMIC / STATISTICAL
UNKNOWN	UNKNOWN	DEEP LEARNING DATA INTENSIVE GENERIC SOLUTIONS

PROBLEM SOLVING → PROBLEM INDEPENDENT

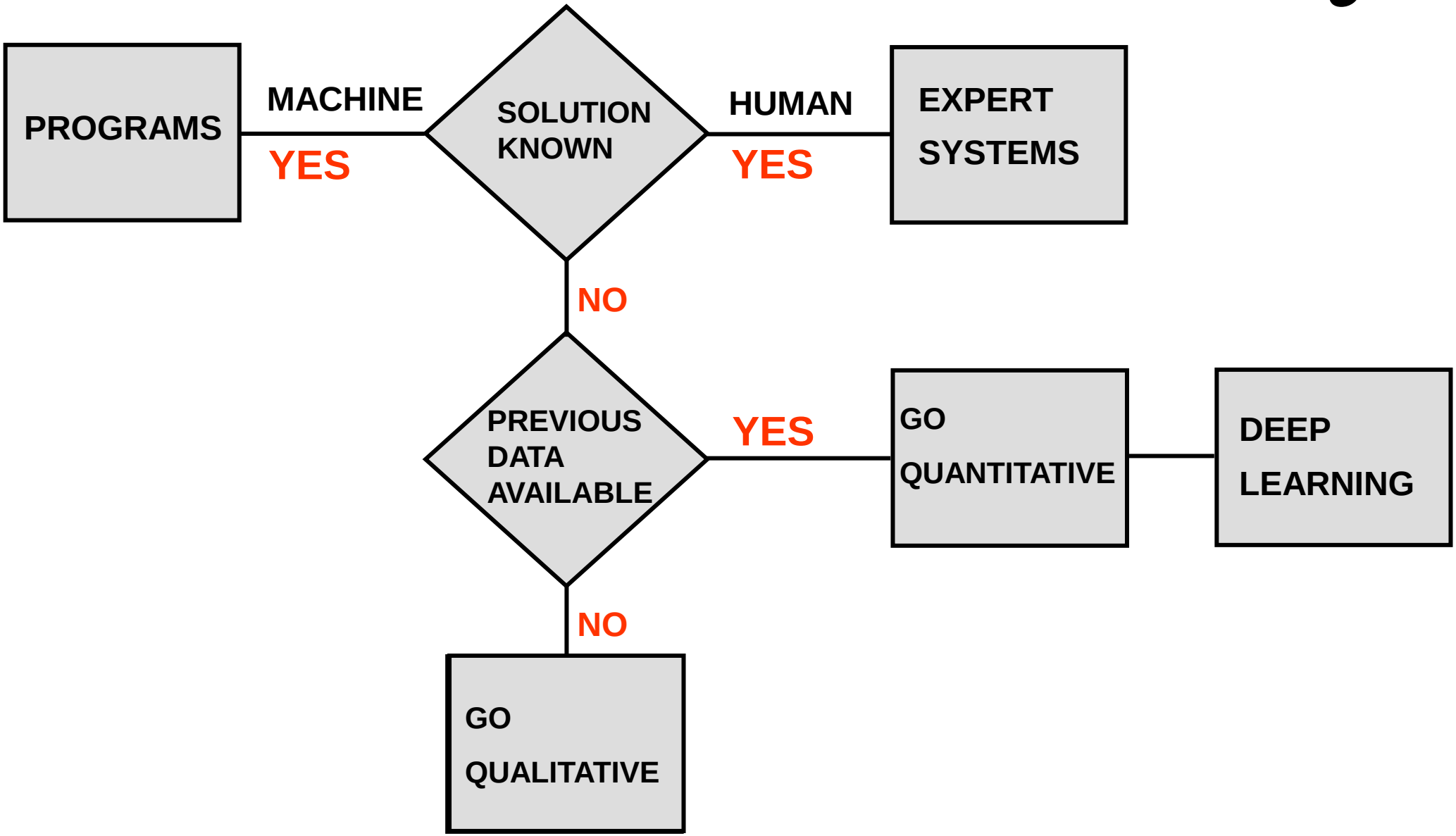
وداعا للتخصص البغيض

يا أهل حوسبة اللغة فلتسهموا في محاربة التخصص المنغلق

أداتكم البيانات وسلاحكم اللغة وزادكم الذي لا ينضب هو معرفة الإنسانيات

PROBLEM SOLVING

حـــل المشكـــلات



لا حلول مع اليأس ولا يأس مع الحلول

QUALITATIVE METHODS

- **NARRATIVE**
- **PHEMONELOGICAL**
- **GROUND THEORY**