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 uns
- **Example:**

ى الاسكندريه 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.

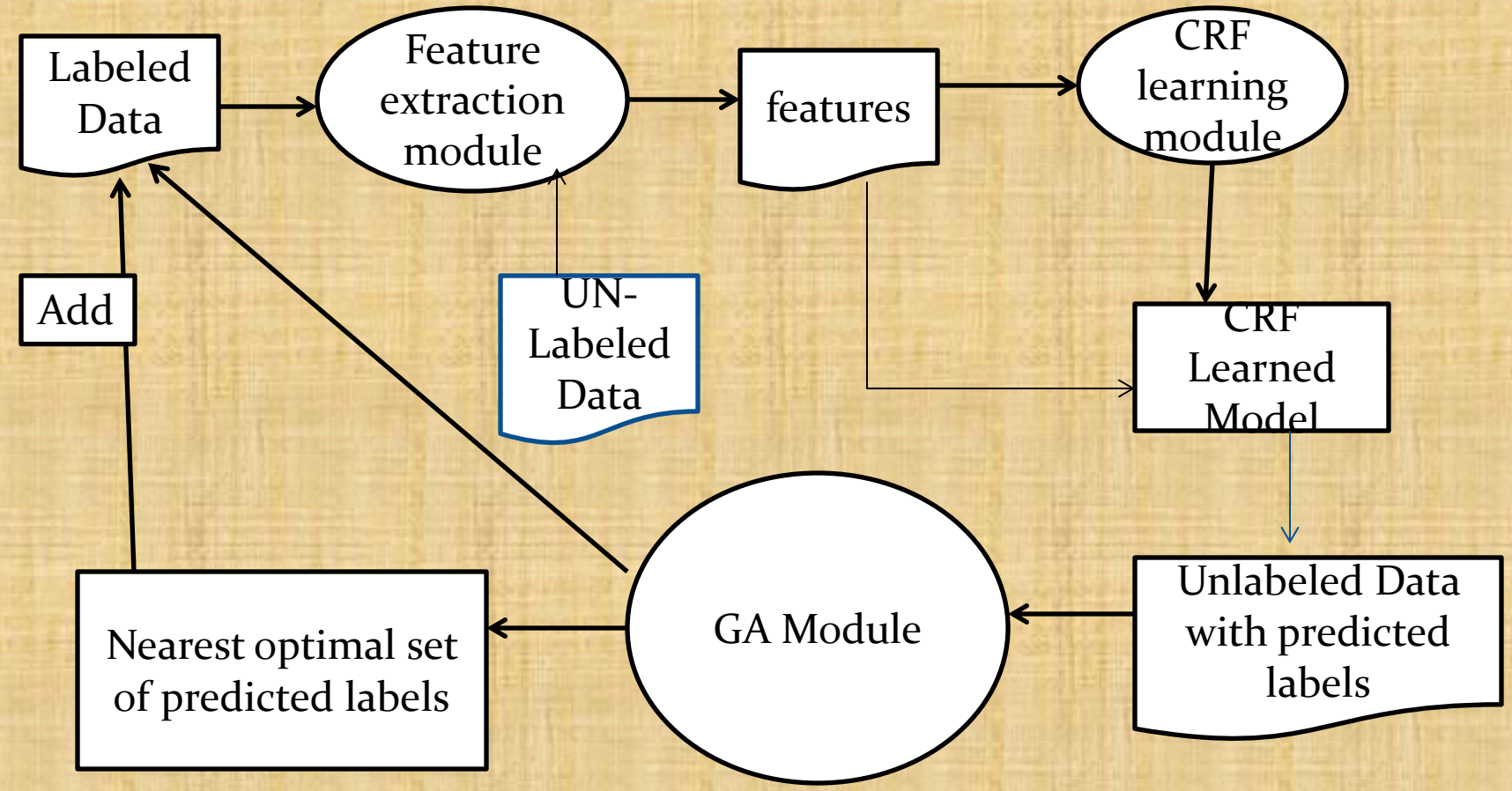
Motivation

- 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')

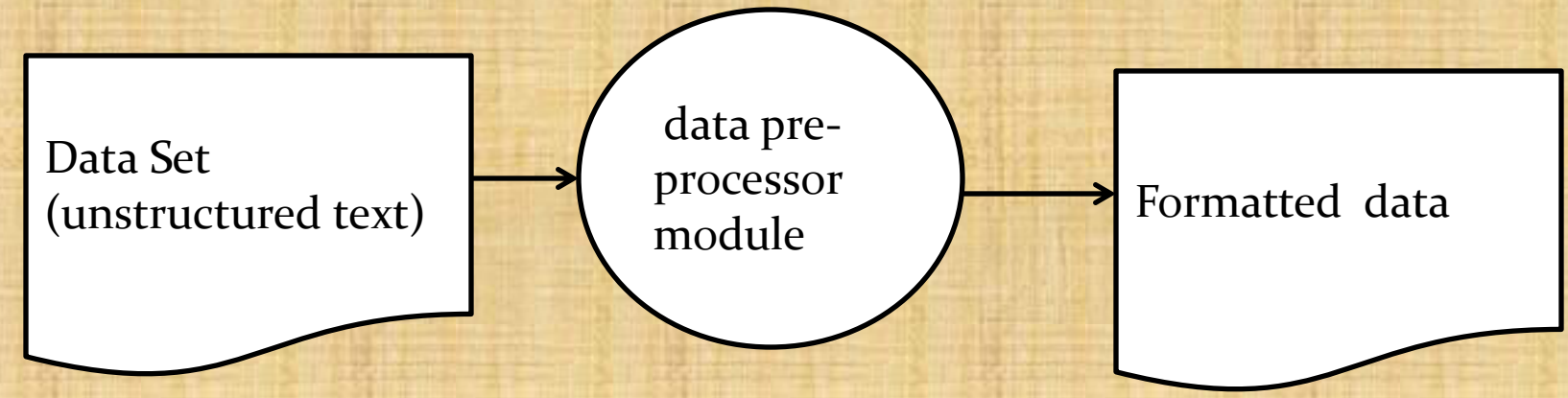
- No other researches have used this combination in semi-supervised learning in ANER field.
- The rational of using only the three types is because the main objective is to try the hybrid algorithm CRF and GA on the basic three types if it achieves good results it will be expand to cover more types of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

Proposed Solution Components

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

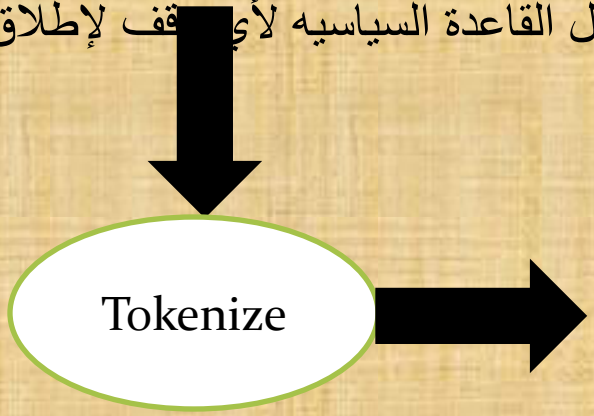
Data pre-processor Module

- 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 السياسيه لأي وقف لإطلاق النار }



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

Data pre-processing module

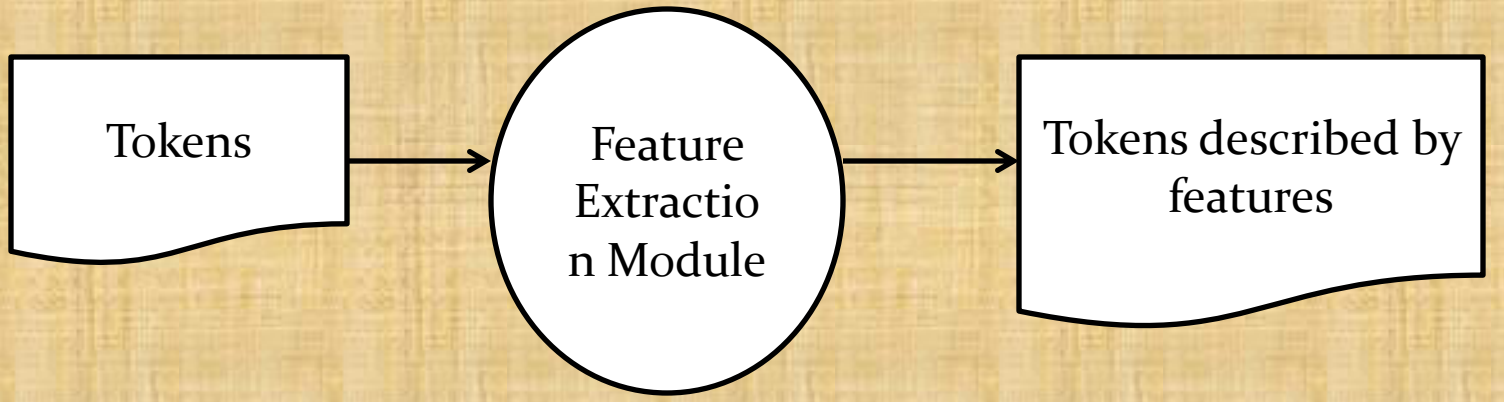
- Data cleaning
  - Data set contains some useless special characters, these special characters removed to make data more reasonable.
  - Such as :
  - “”
  -)
  - +,-,.,etc

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

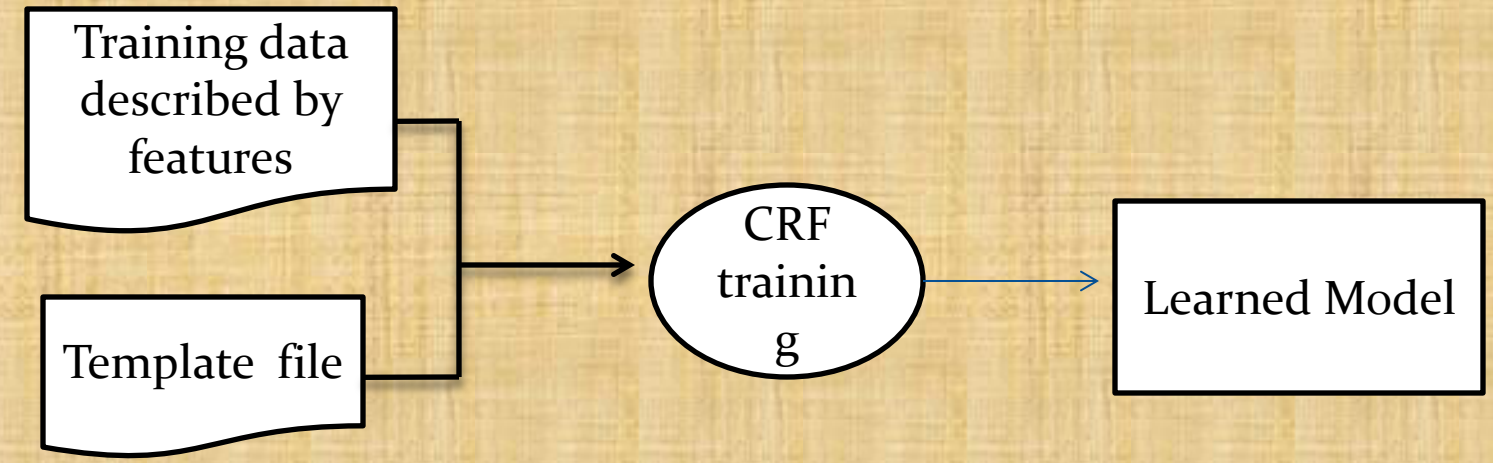
```
token person_gaz org_gz loc_gaz pers_Ind Org_ind loc_ind
اعلن false false false true false false O
اتحاد false false false false true false B-ORG
صناعة false false false false false false I-ORG
السيارات false false false false false false I-ORG
في false false false false false false O
المانيا false 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
```

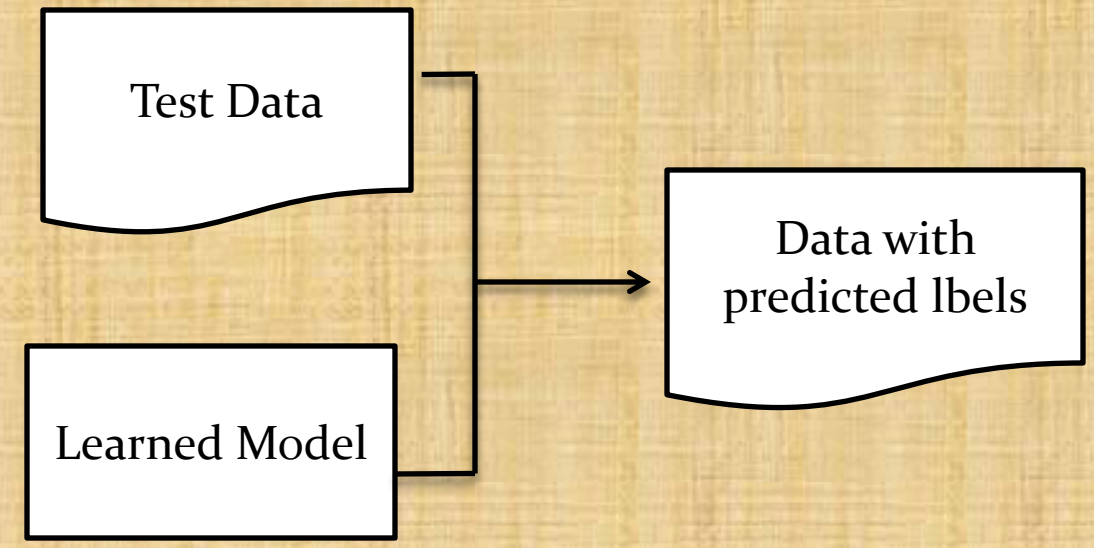
CRF Module

- CRF training module



CRF Module

- CRF Test



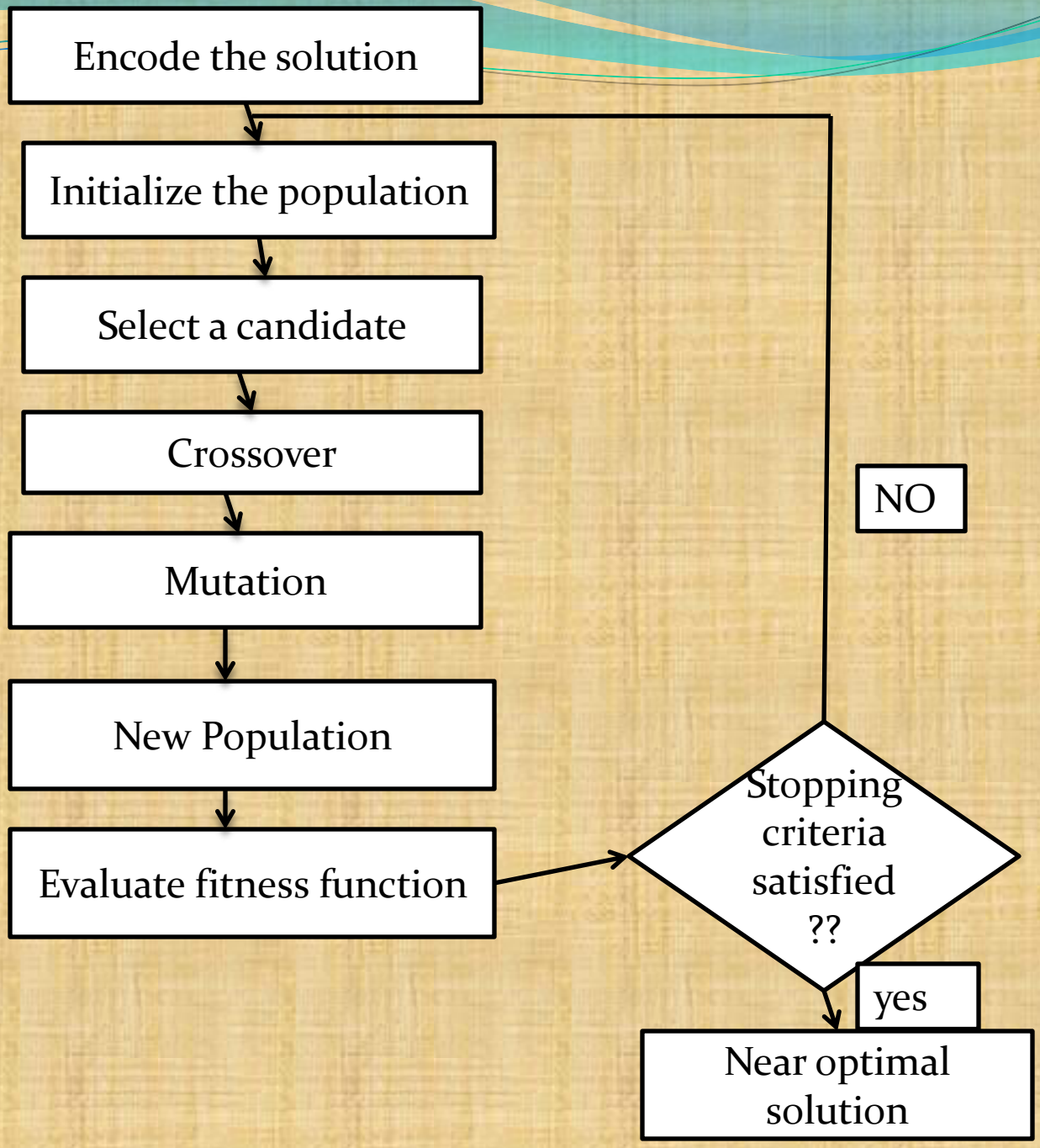
Tokens with Predicted labels

```
# 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
الاقاد 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
برند 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
جونشولك 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
للاقاد 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

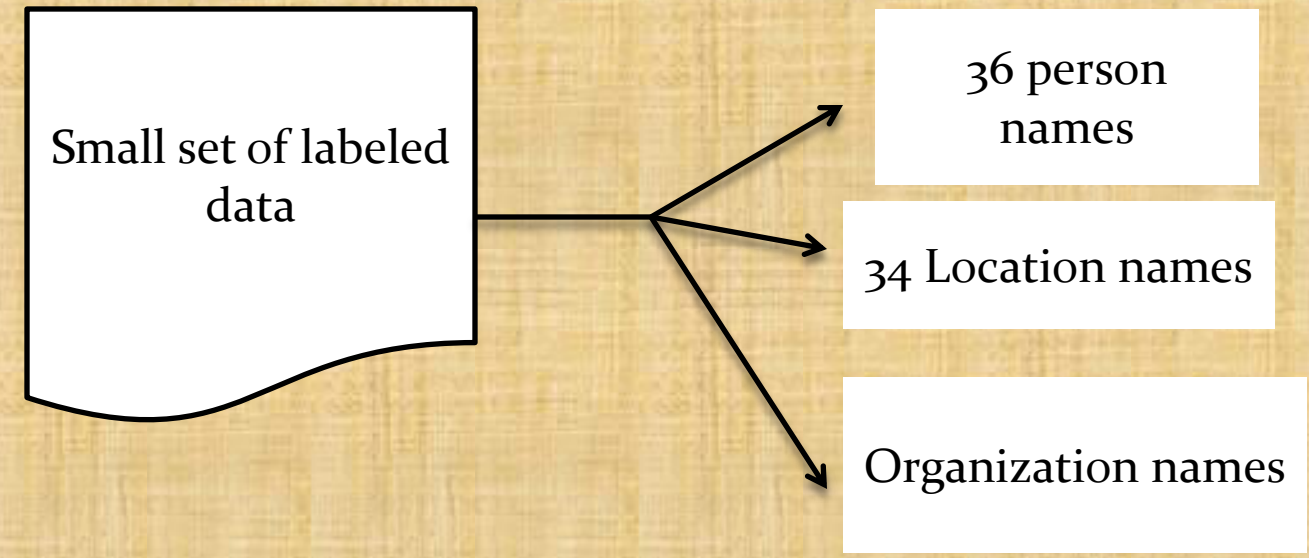
Encoding the solution

- 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’)

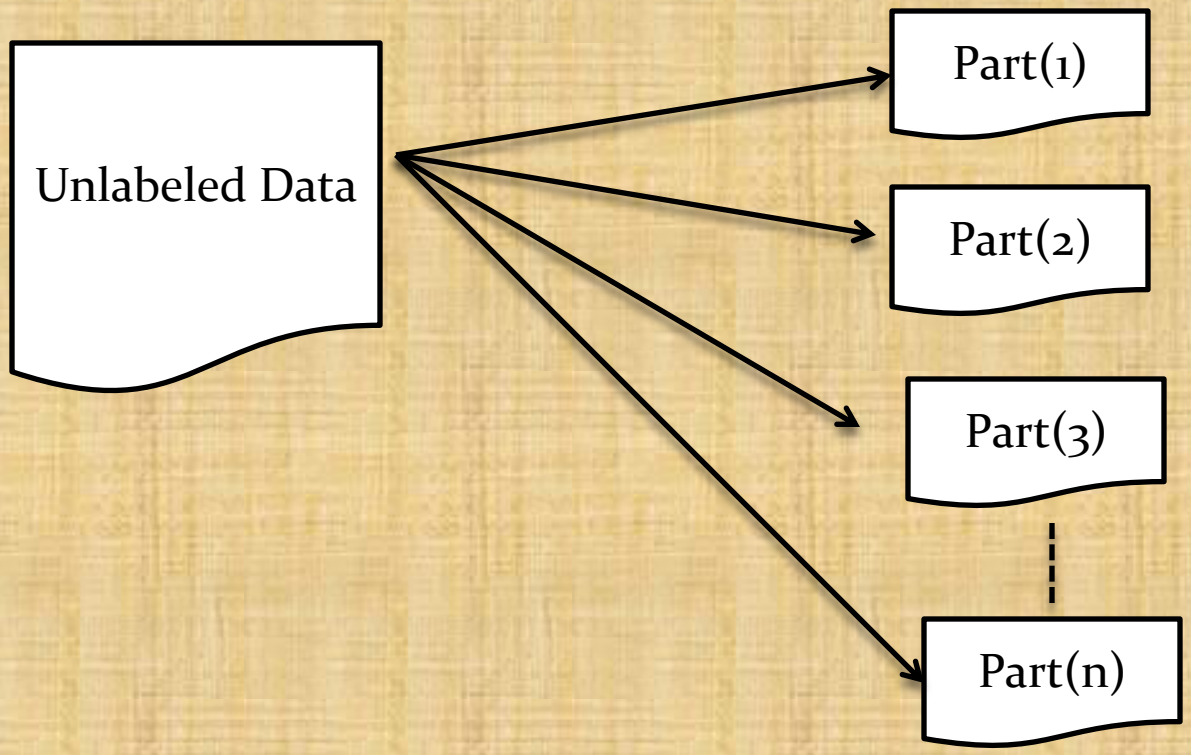
- As mentioned before ,the objective of GA is to add unlabelled data with predicted labels to the labeled training data to maximize its size and therefore enhance the accuracy of the system

Initial set of labeled data(Seeds)

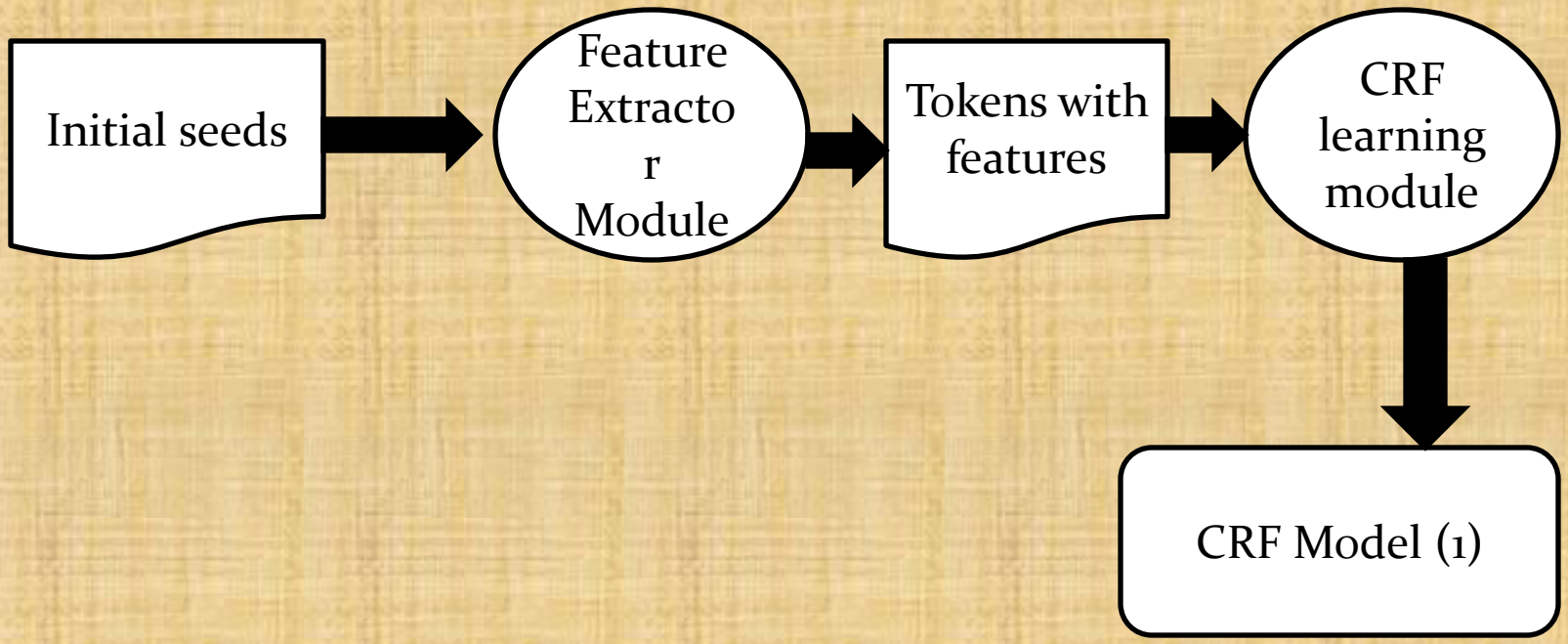


Unlabeled Data

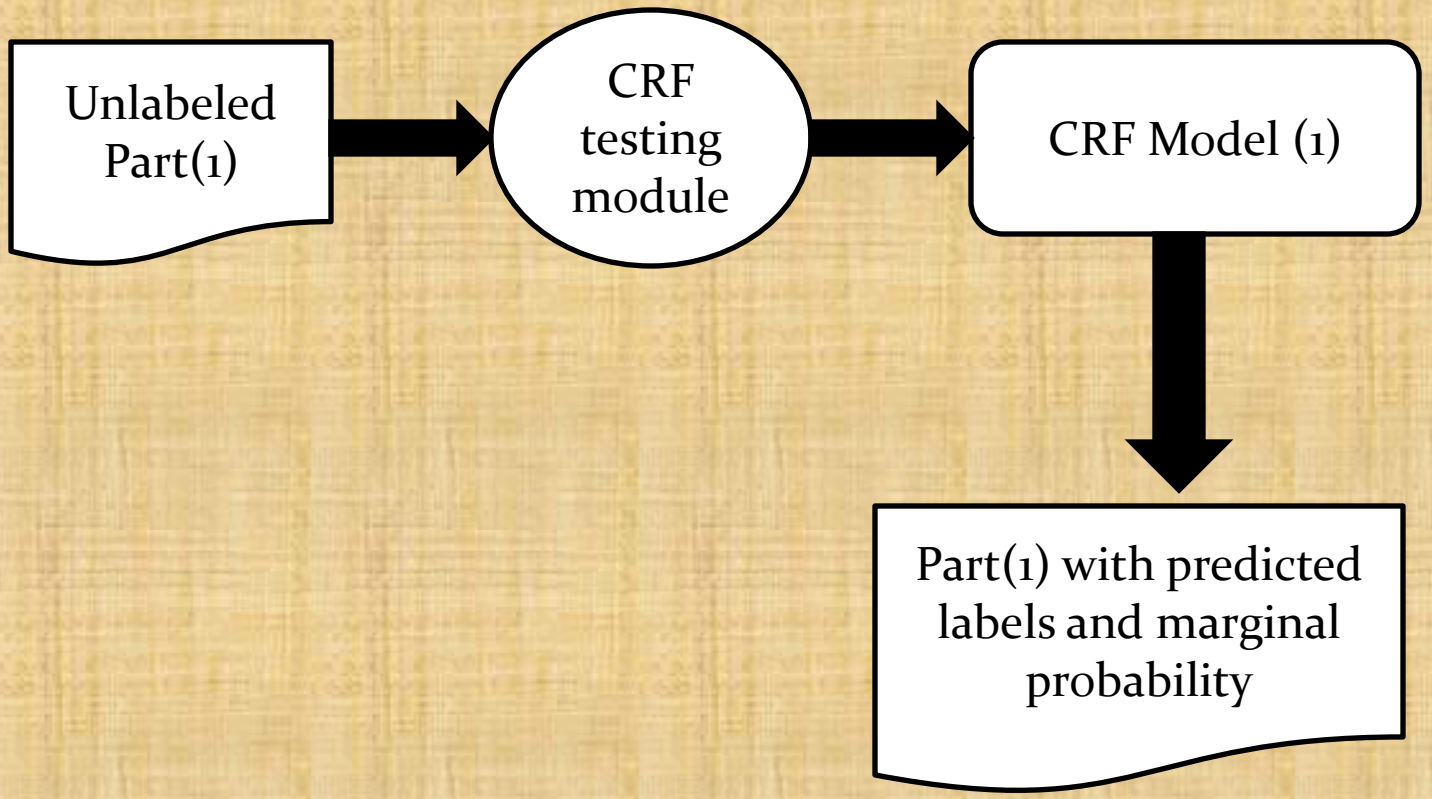
- Unlabeled data is partitioned into some parts each part is about 400 tokens

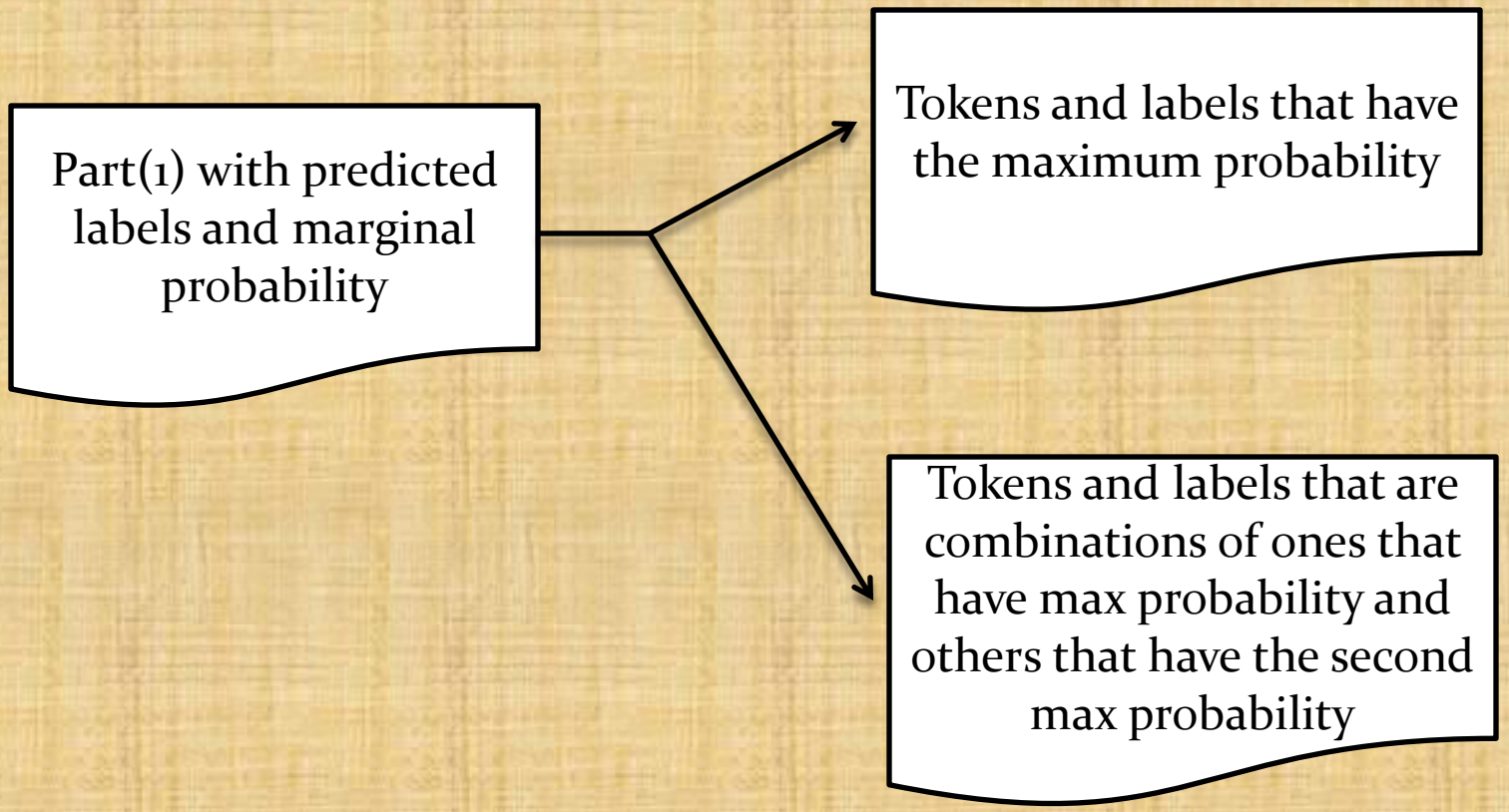


Iterations and unlabeled data addition

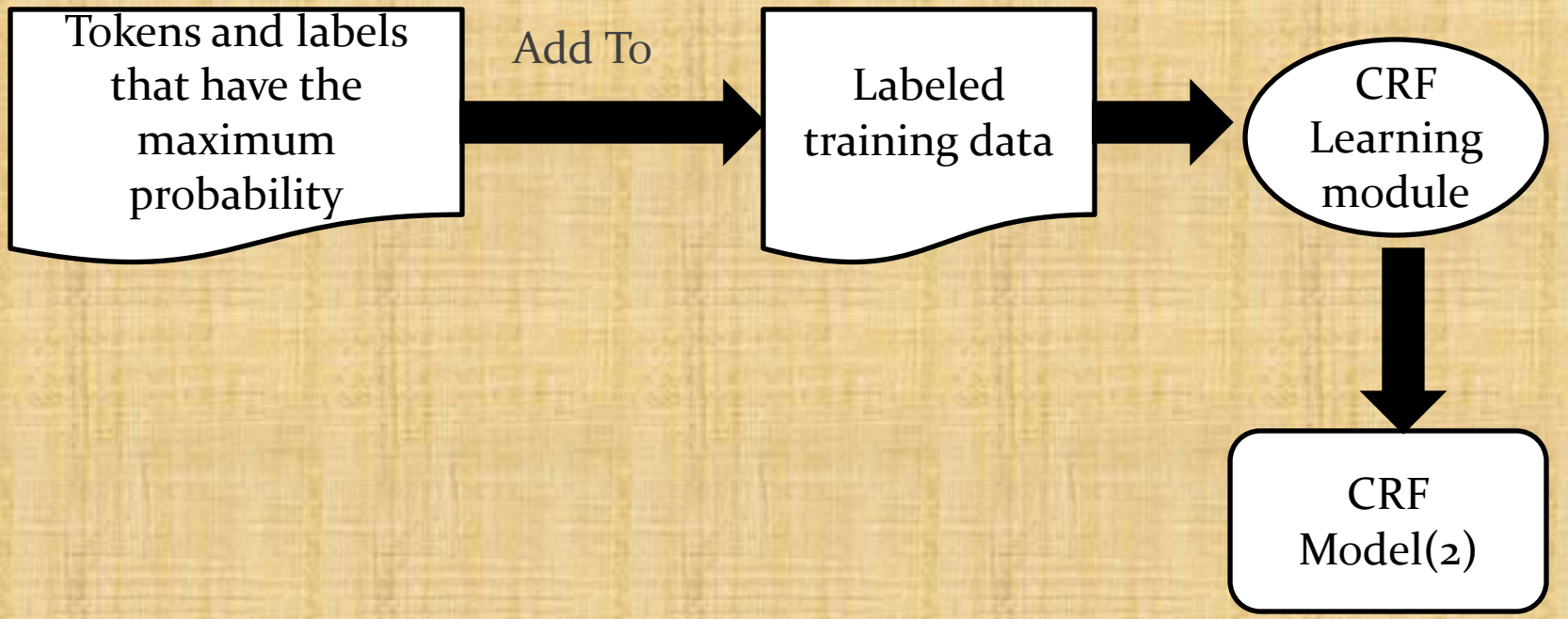


Iterations and unlabeled data addition



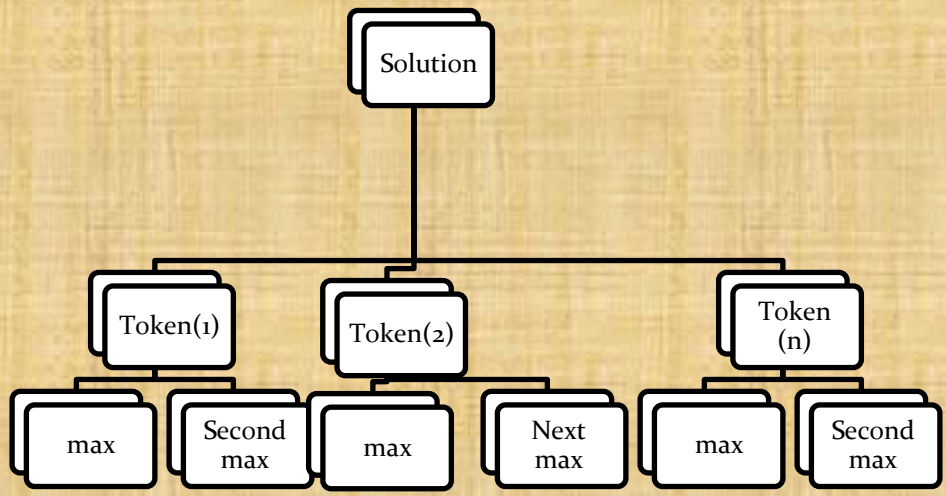

Part(1) with predicted labels and marginal probability
Tokens and labels that have the maximum probability
Tokens and labels that are combinations of ones that have max probability and others that have the second max probability

Using only CRF



Using both CRF and GA

Tokens with predicted labels and probabilities



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 :

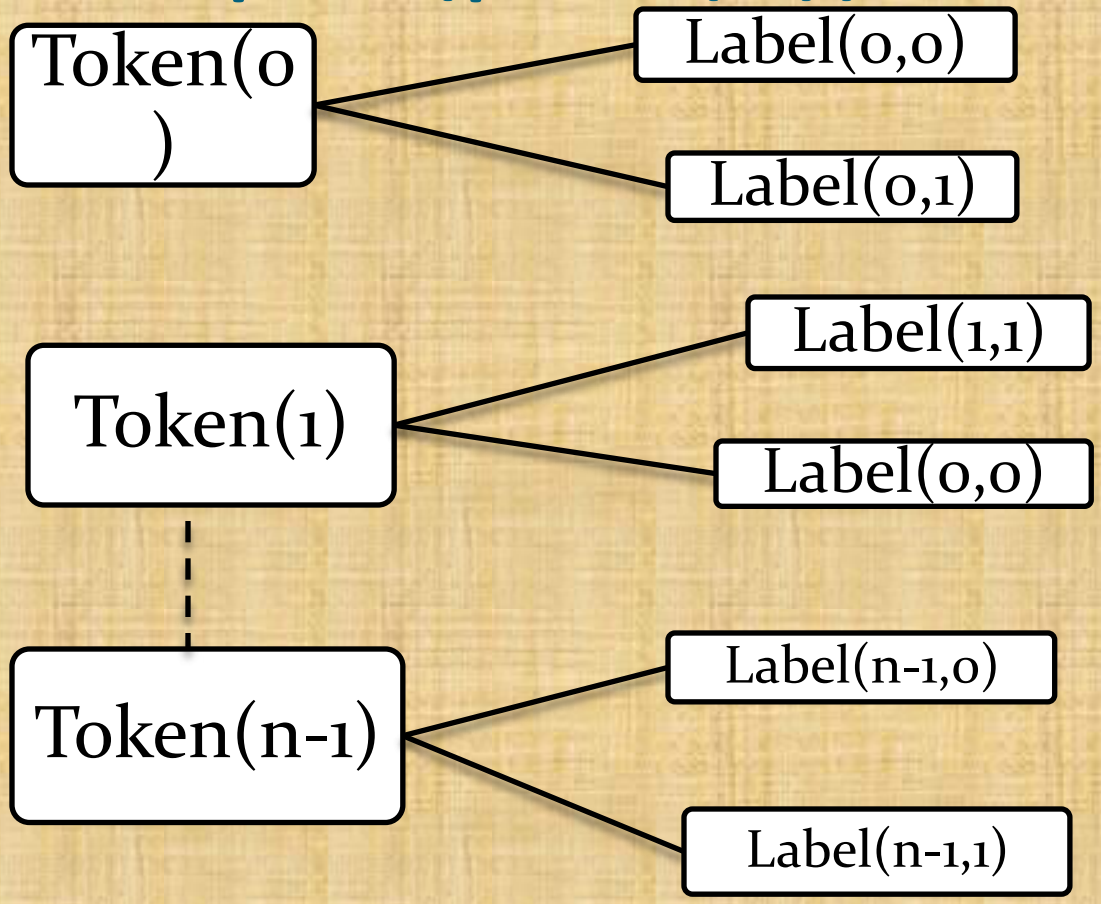
  0→the token takes the label with max probability

  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

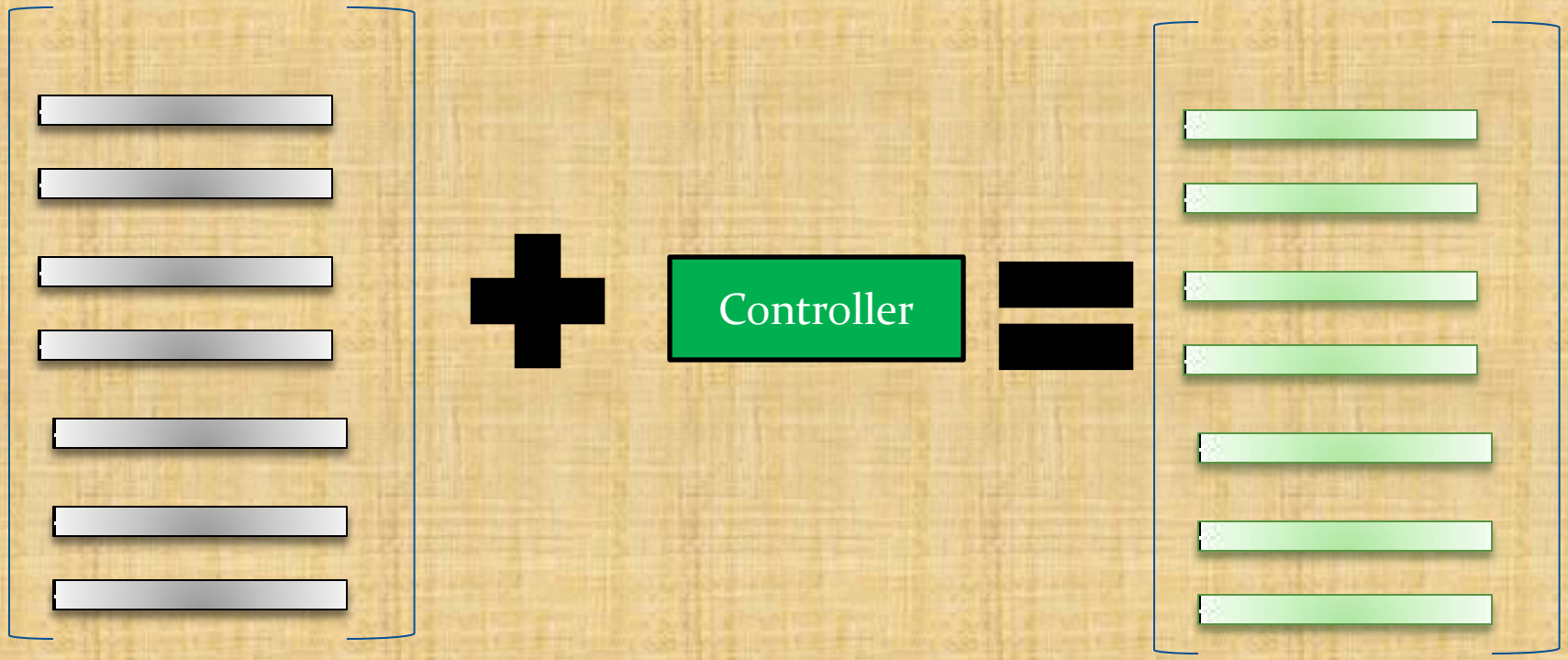
population

- 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)

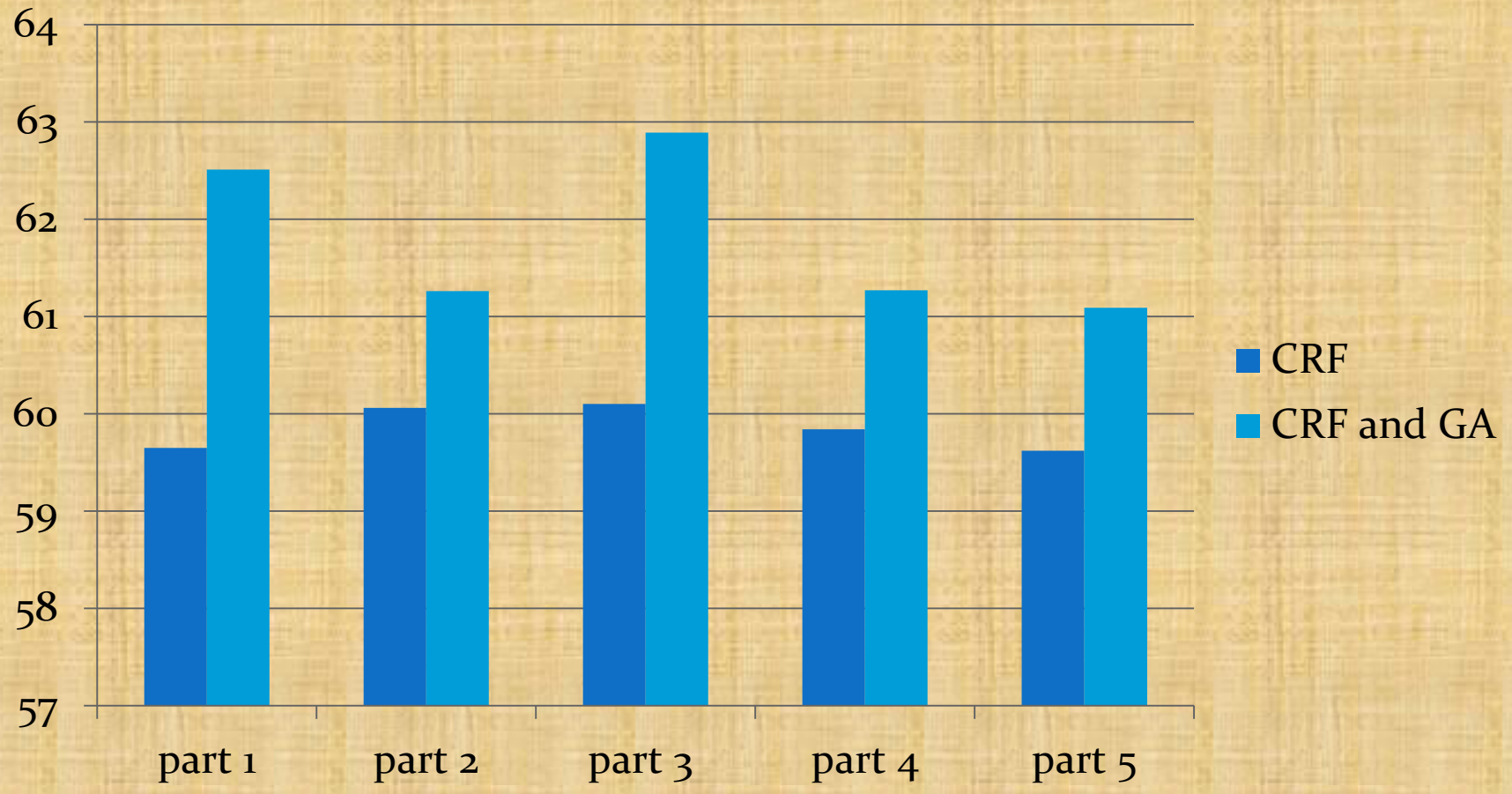
Using only CRF

	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!

هنسة عين شمس

جمعية هندسة اللغة

2013-12-11

فـروع الإنسانيـــات الجـدد وحوسبـــة اللغـــة

د. نبيل علي

INSPIRE OR EXPIRE

الإطـــار العـــام

1 • الإنسانيات : النقلة النوعية

2 • موسم الهجرة إلى الجمعي

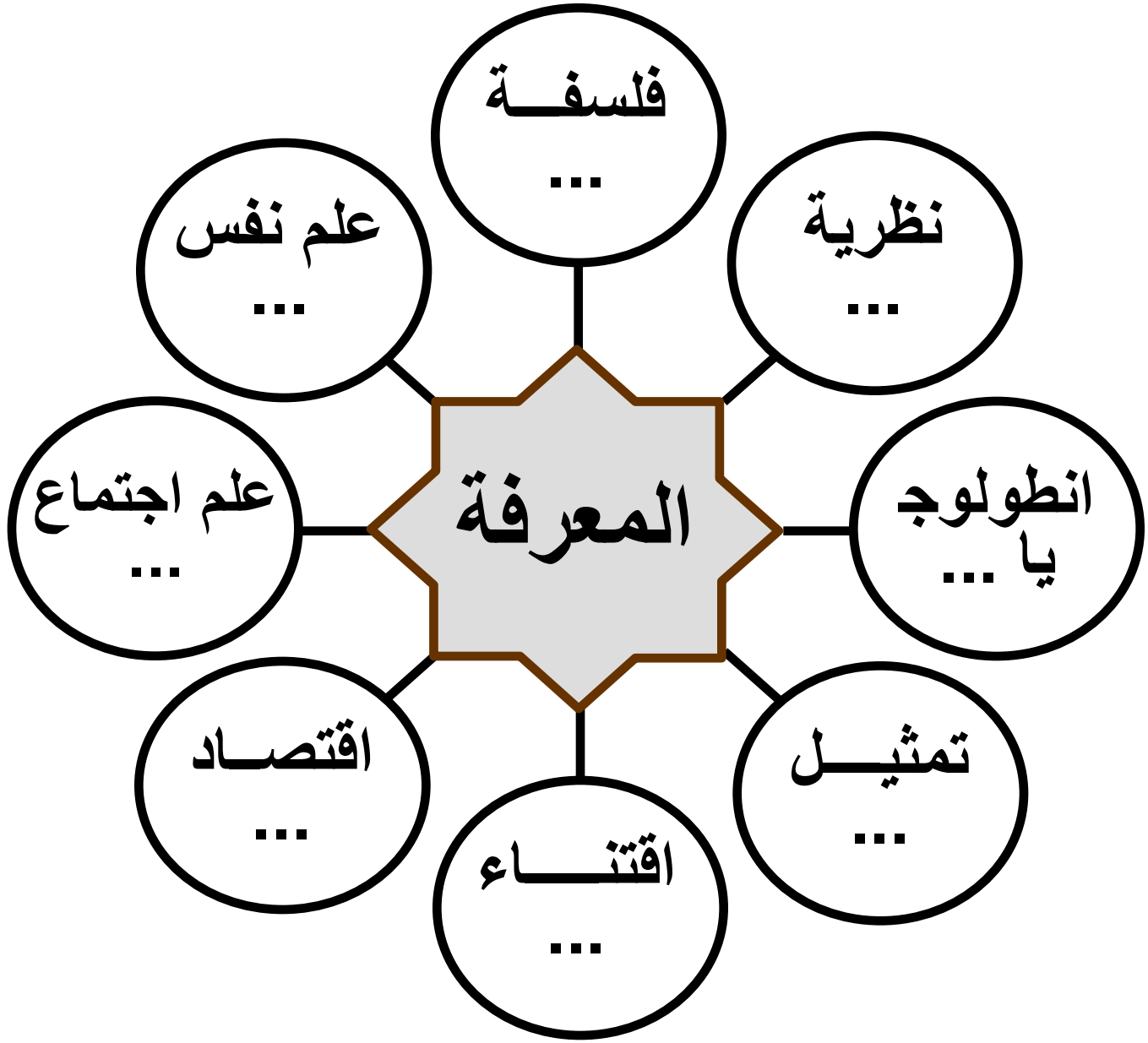
3 • حوار الشبكات

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

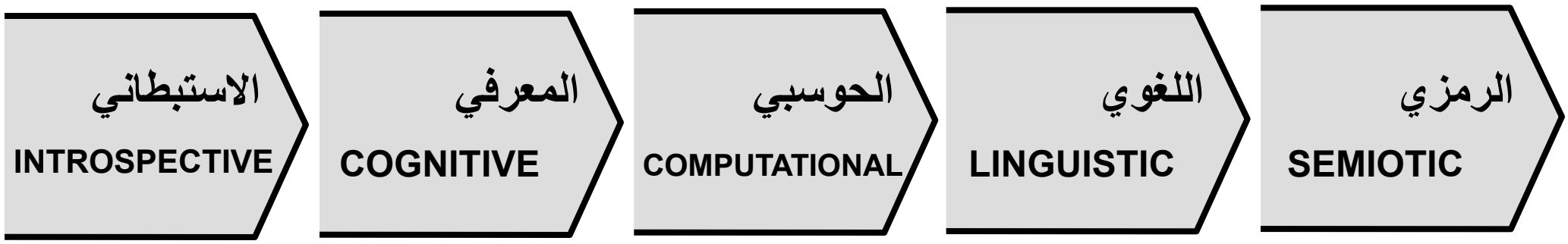
دعنا ننفض الضوضاء عن الظاهر المخادع الزائغ للكشف عن النظام الكامن في جوفه

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

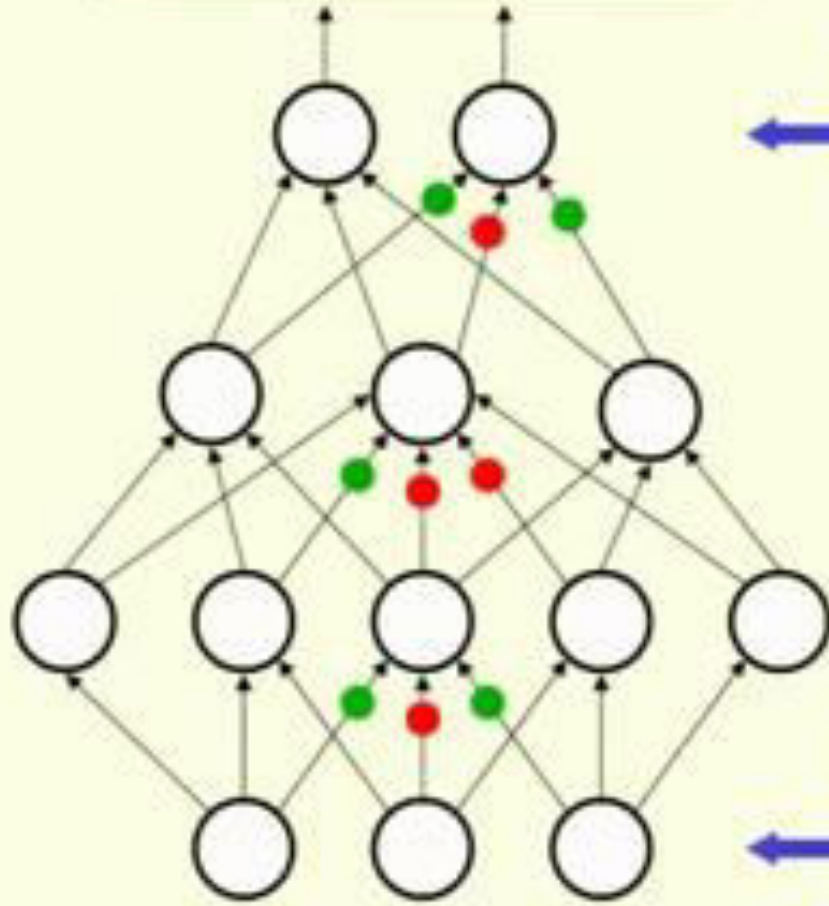
CLUSTERING

...

Deep neural networks (~1985)

Back-propagate error signal to get derivatives for learning

Compare outputs with correct answer to get error signal



outputs

hidden layers

input vector

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

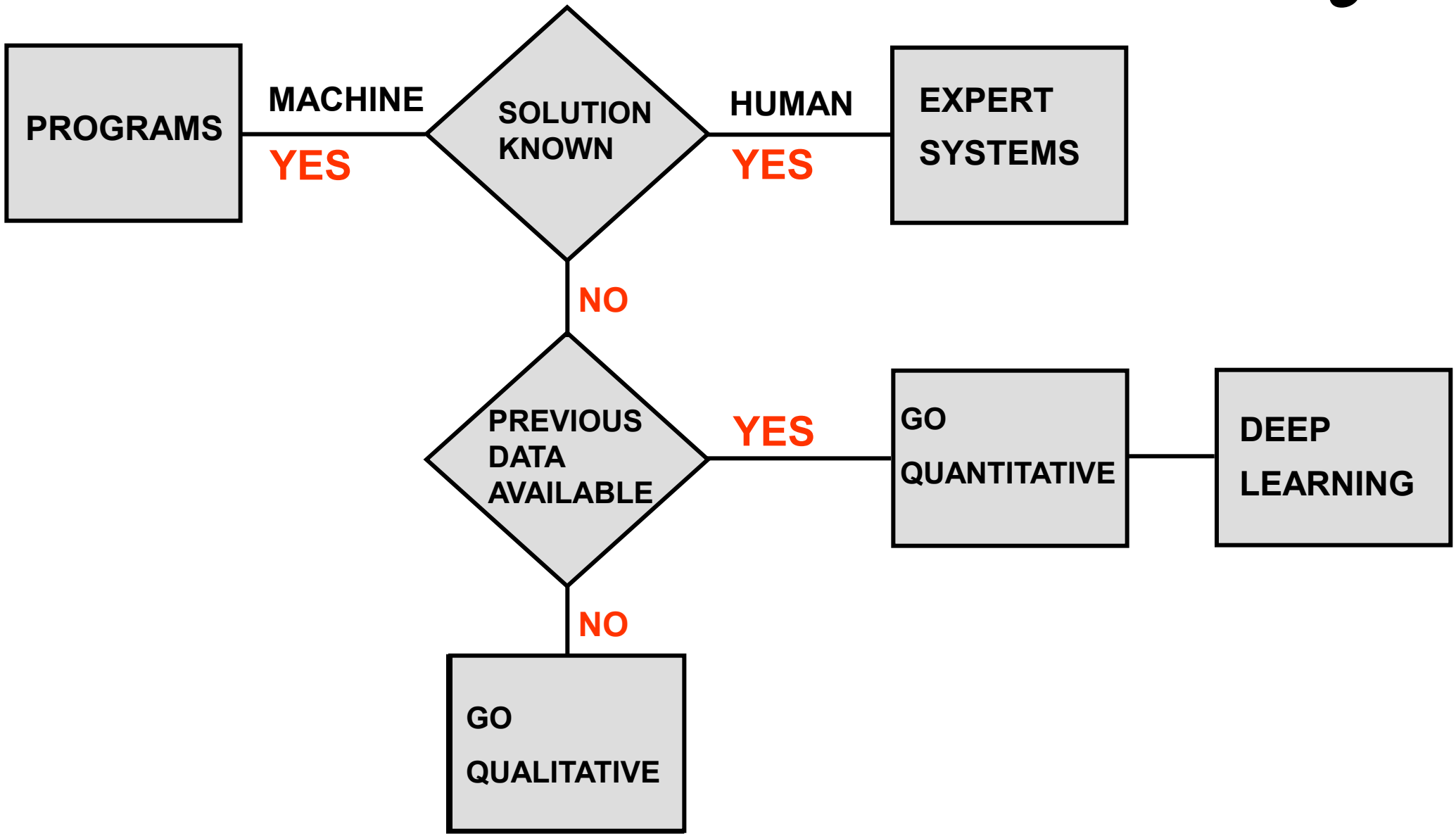
وداعا للتخصص البغيض

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**

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]

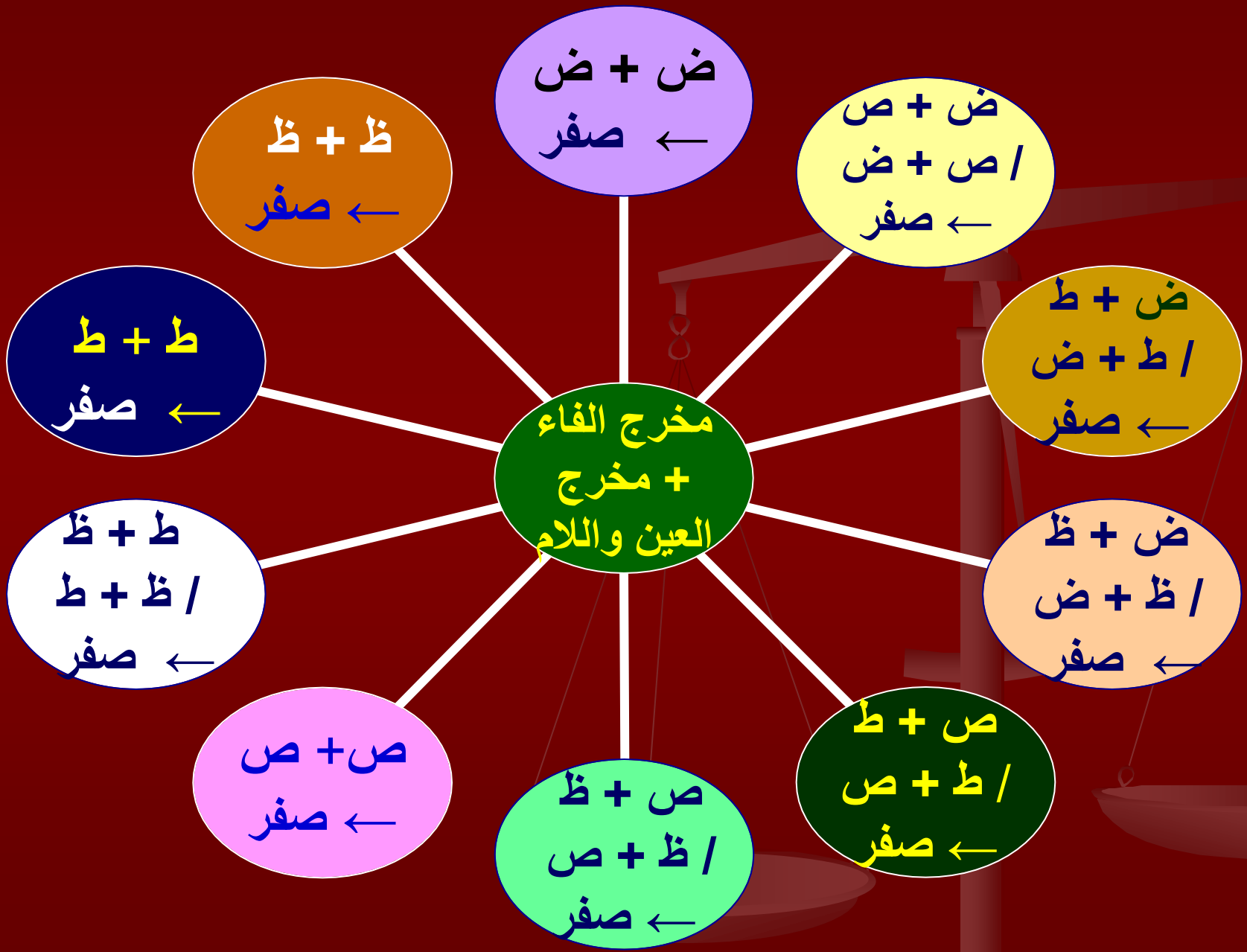
- الأصوات الشفهية : (ف - ب - م) [labial]

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	-	-			-		-	-	-		-	-		-	-	-	-	-	-	-						
تاء	-	-		-	-		-		-	-	-	-	-	-	-	-	-	-	-	-			-	-		
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

5/1/2024

Agenda

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

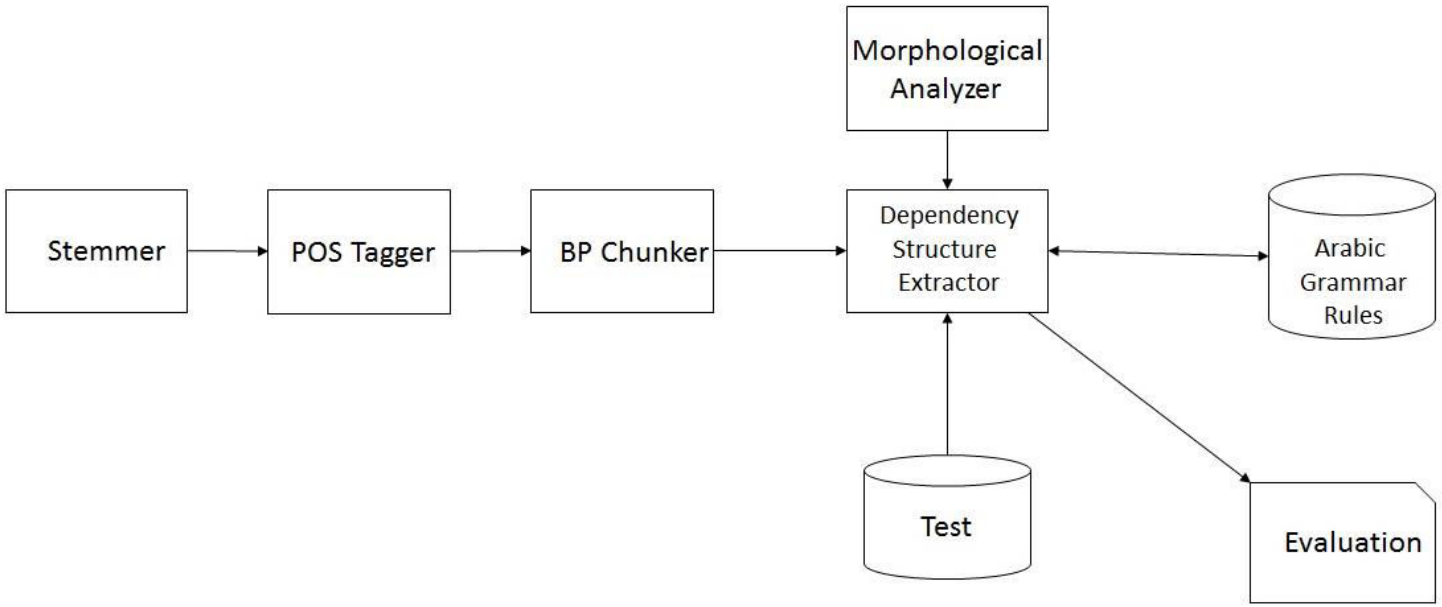
Problem Definition

- 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.

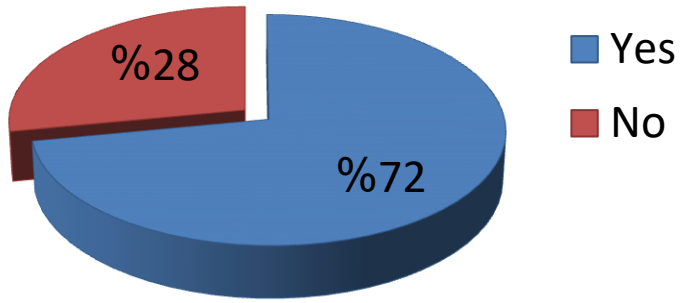
Introduction

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

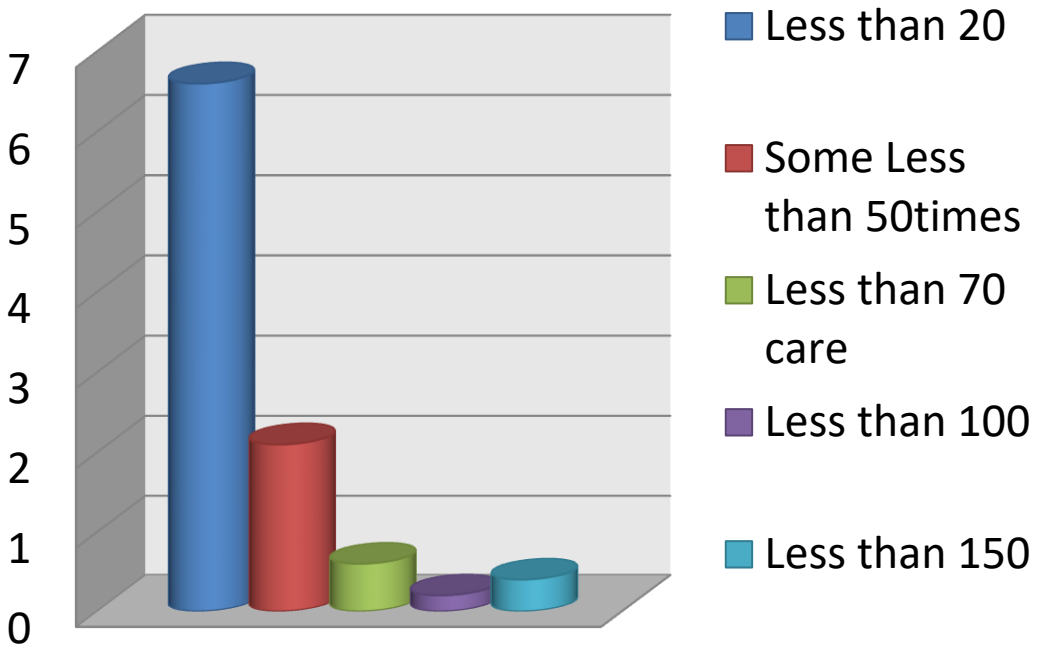
رصد	4,197,917 likes
الصفحه الرسميه لرئاسه مجلس الوزراء	1,294,174 likes
شبكه اخبار مصر	1,174,196 likes

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

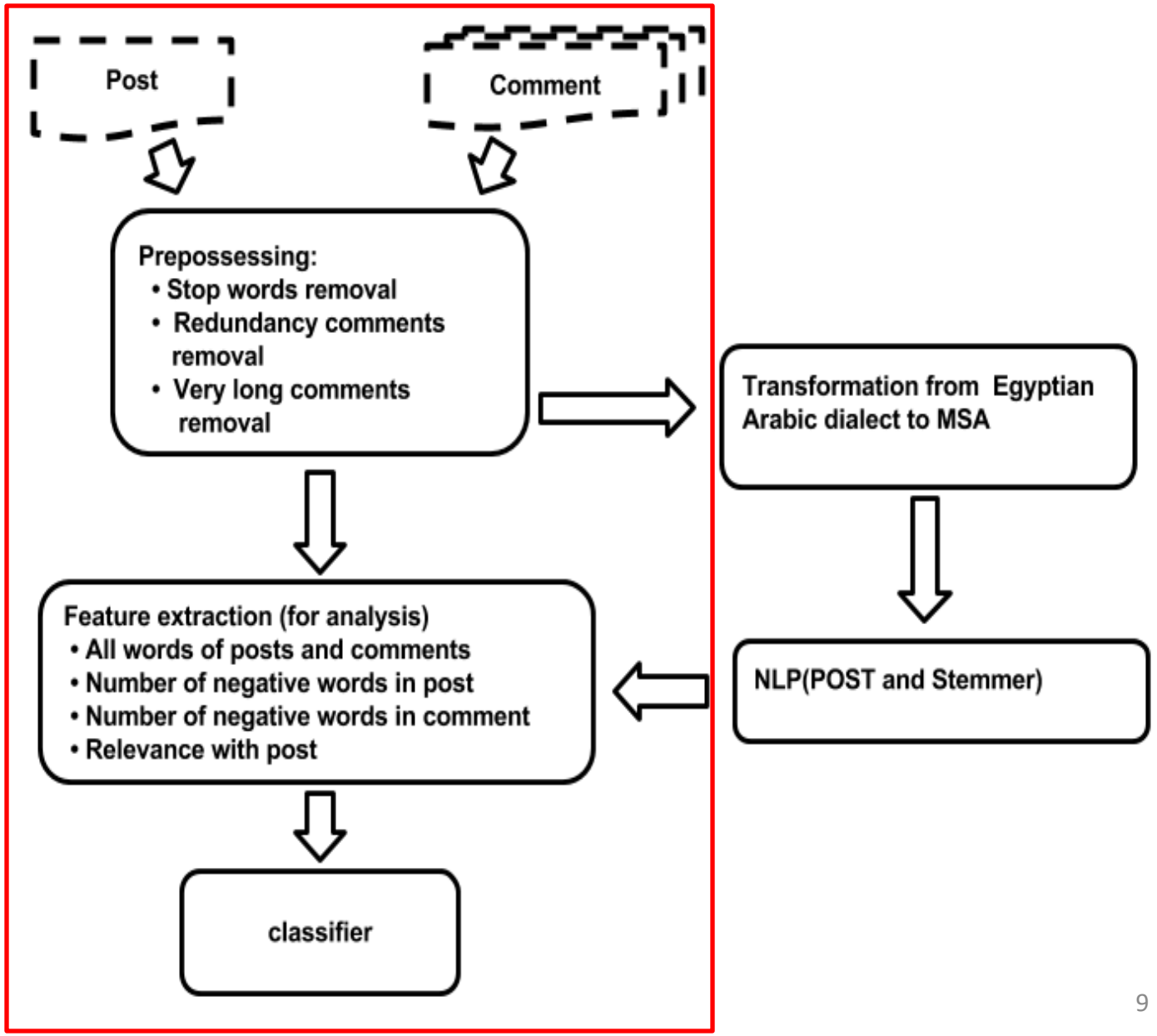
75 people like this.

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

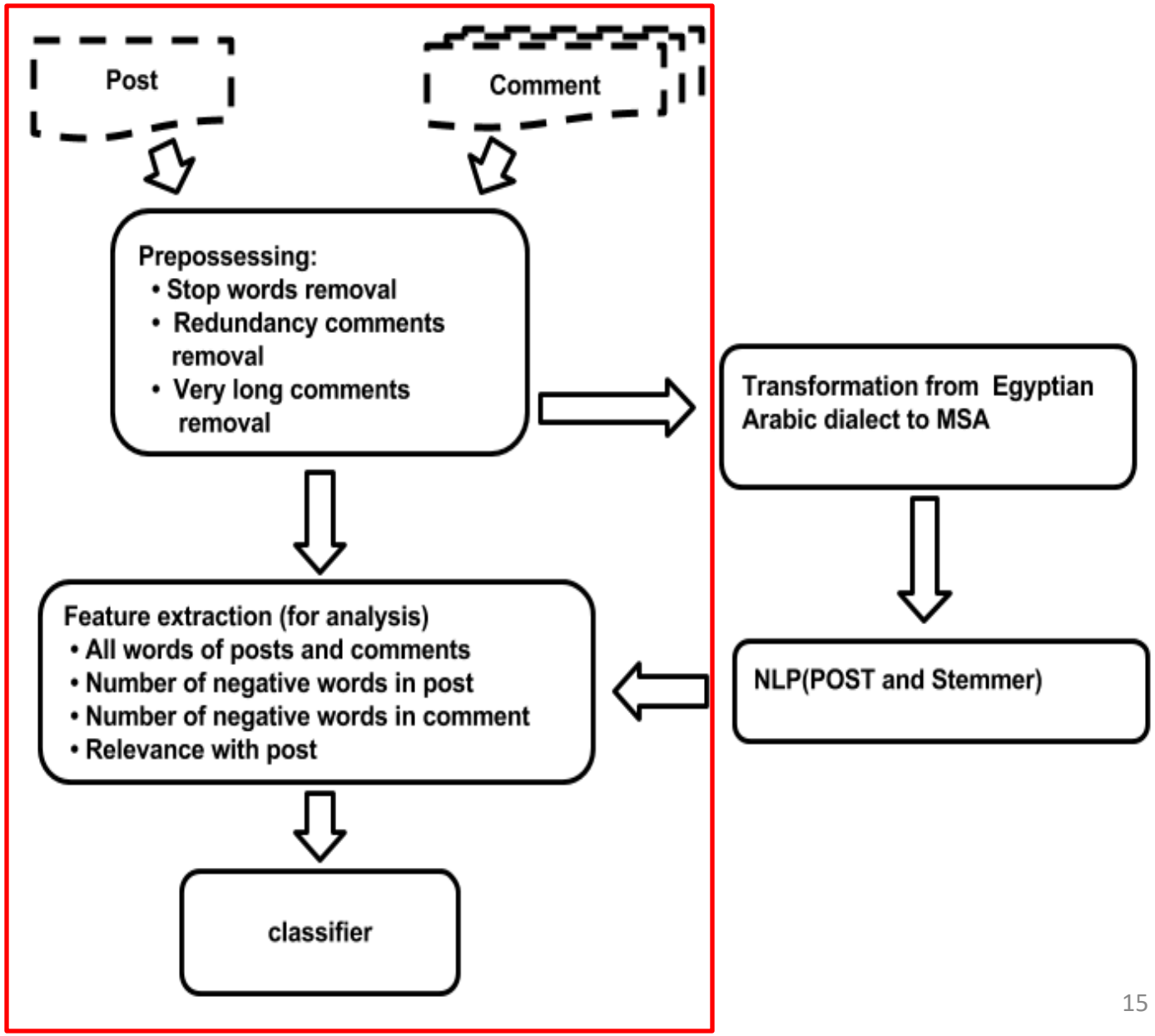
4. Relevance with post

$$Tf = F\ (1)$$

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

Then we apply SVM on these features.

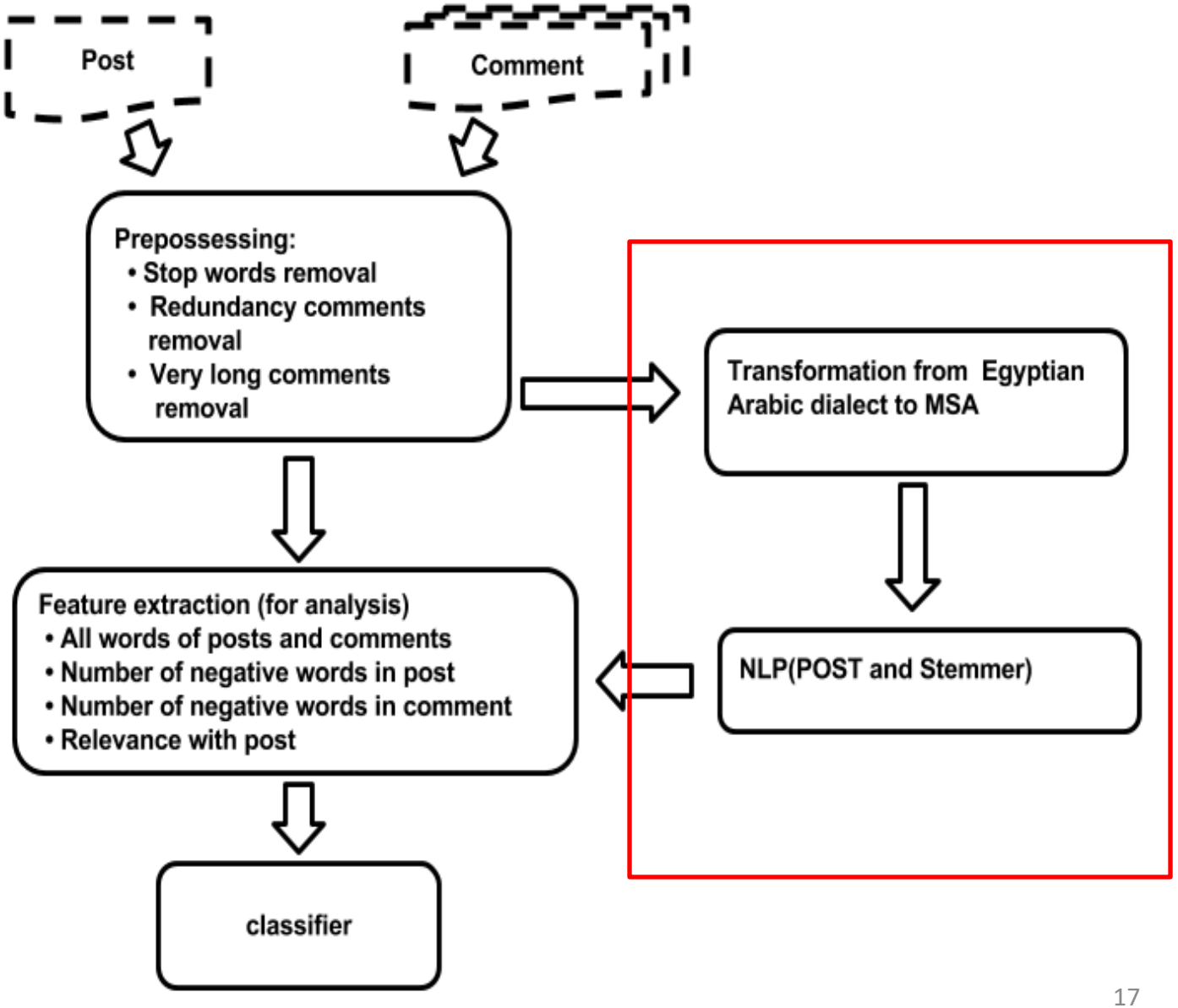
Sentiment Analyzer Block Diagram



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

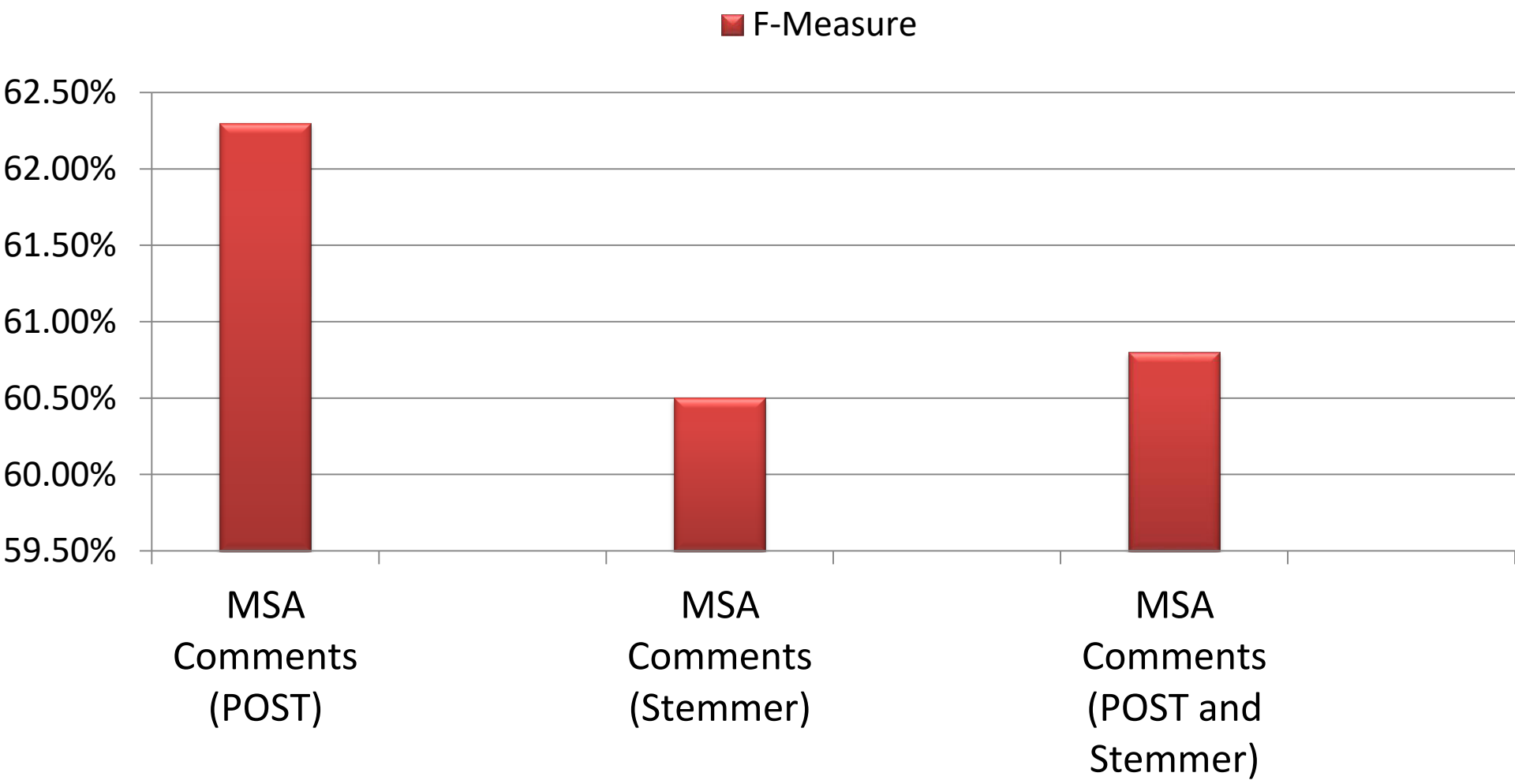
stemmer takes an Arabic word and returns the stem of it. In this case, the stem of the word is used as a feature. So we replace a word "يلتقى" with "لقي".

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