Arabic Morphological Analyzer

GRADUATION PROJECT

Supervisor:

- د. عزة طه

Team Group:

- أحمد خالد محمد كامل
- محمد عبد الوهاب سيد
- أحمد محمد محمود
- يوسف محمد سعيد

[إحصاء حاسب]

Project Members

Agenda

- What is a Morphological Analyzer? And why it is important?
- Main Approaches to build a Morphological Analyzer
- Morphological Analysis in Arabic language
- Project Stages
- Additional Features
- Challenges of Arabic Morphological Analyzer
- Incomplete tasks
- References

What is a Morphological Analyzer?

- A morphological analyzer is a tool or software program that is used to analyze the structure of words in a language, specifically their morphology. Morphology is the study of how words are formed and how they relate to one another.
- A morphological analyzer typically takes a word as input and then breaks it down into its constituent morphemes, which are the smallest units of meaning in a language. It then provides information on the grammatical features of the word, such as its part of speech, tense, gender, and number.

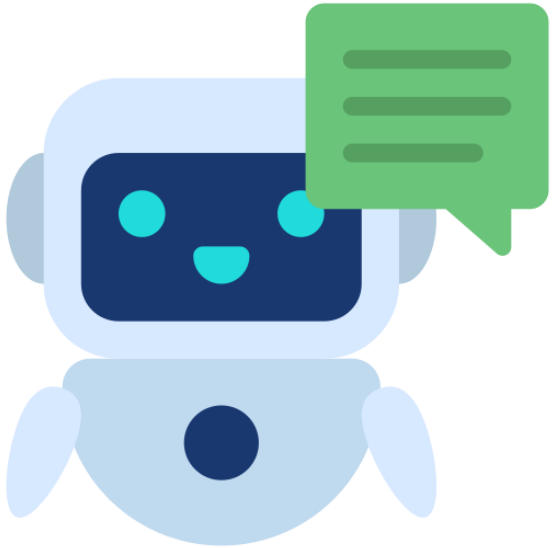
Played ➔ Play

كتب ➔ يكتبون

Why it is important?

- Morphological analyzers are used in a variety of applications in natural language processing (**NLP**) such as **machine translation** and **Information retrieval**.
  They can be especially useful in languages with complex morphology, where words may have many different forms depending on their context and the grammatical rules of the language.

- Overall, a morphological analyzer is an important tool for anyone working with a language, as it can help in understanding the structure of words and how they are used in context.



Why we need root of words ?

- A root word is the most basic form of a word.
- Root words can help you to break down large, new words into smaller units to discover their meanings.
- Learning just one root word can help you understand several words.

Main Approaches to build a Morphological Analyzer

DATA DRIVEN APPROACH

RULE-BASED APPROACH

Main Approaches to build a Morphological Analyzer

Data Driven Approach [ML]

- A data-driven morphological analyzer is a type of morphological analyzer that uses large amounts of data to learn the rules and patterns of word formation in a language. This contrasts with rule-based morphological analyzers, which rely on pre-defined rules to analyze words.

- With a data-driven approach, the morphological analyzer is trained on a large corpus of text, which is typically annotated with information about the morphemes and grammatical features of each word. The analyzer then uses machine learning algorithms to learn the patterns and rules of word formation based on the data.

- The advantage of a data-driven approach is that it can capture the complex and varied patterns of word formation in a language, which may be difficult to capture with pre-defined rules. This can lead to more accurate analyses of words and better performance in natural language processing applications.

Main Approaches to build a Morphological Analyzer (.count)

Rule-Based Approach

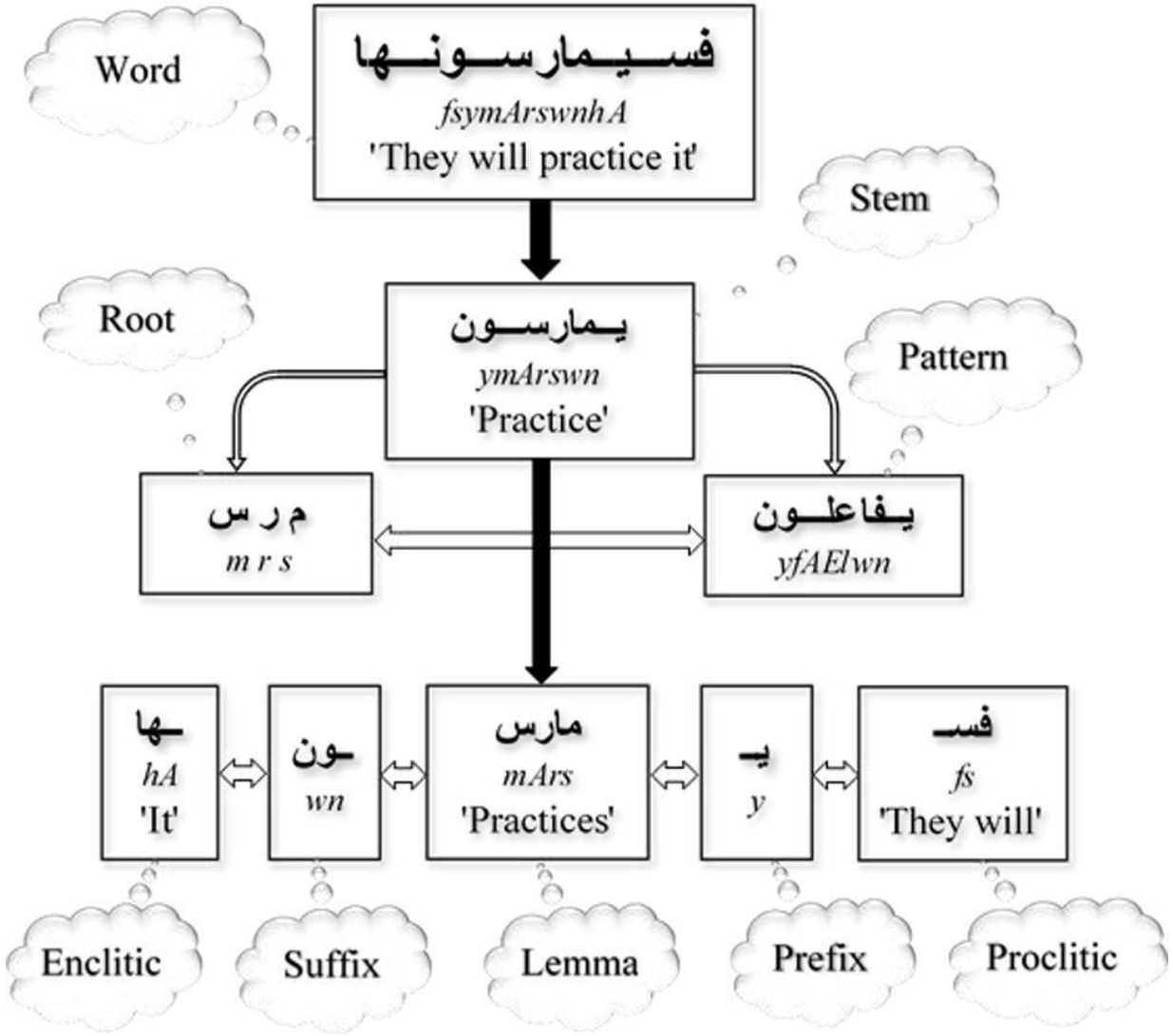
Verb + ed
Ex: played = play + ed

- A rule-based morphological analyzer is a type of morphological analyzer that uses pre-defined rules to analyze the structure of words in a language. These rules are based on the known patterns of word formation in the language and are typically created by linguists or language experts.
- The rules specify how to break down a word into its component morphemes, and how to identify the grammatical features of each morpheme and the word. For example, in English, a rule-based morphological analyzer might use rules to identify the past tense of regular verbs, such as adding "-ed" to the base form of the verb.
- The advantage of a rule-based approach is that it can be more transparent and interpretable than a data-driven approach, since the rules are explicitly defined and can be modified by experts as needed. Rule-based analyzers can also be faster and more efficient than data-driven analyzers, since they do not require large amounts of data to train.

In this Project, We used the **Rule-Based** Technique

Morphological Analysis in Arabic language

Arabic Language Characteristics



Arabic Language Characteristics (.count)

Prefixes

The Prefix is a word part added to the beginning of a word to create a new meaning.
In Arabic, there are many letters that can be in the beginning of the word { أ, ال, ب, س, ف, ك, ل, و}

Suffixes

The Suffix is a word part added to the end of a word to create a new meaning.
There are many letters that can be in the end of the word { ا , ات , ان , ة , ت ,تم , ك , كم , ن , ه , هم , و , وا , ون , ي , ين}

Therefore, morphological analyzer should delete these additions to come with the root of the word.

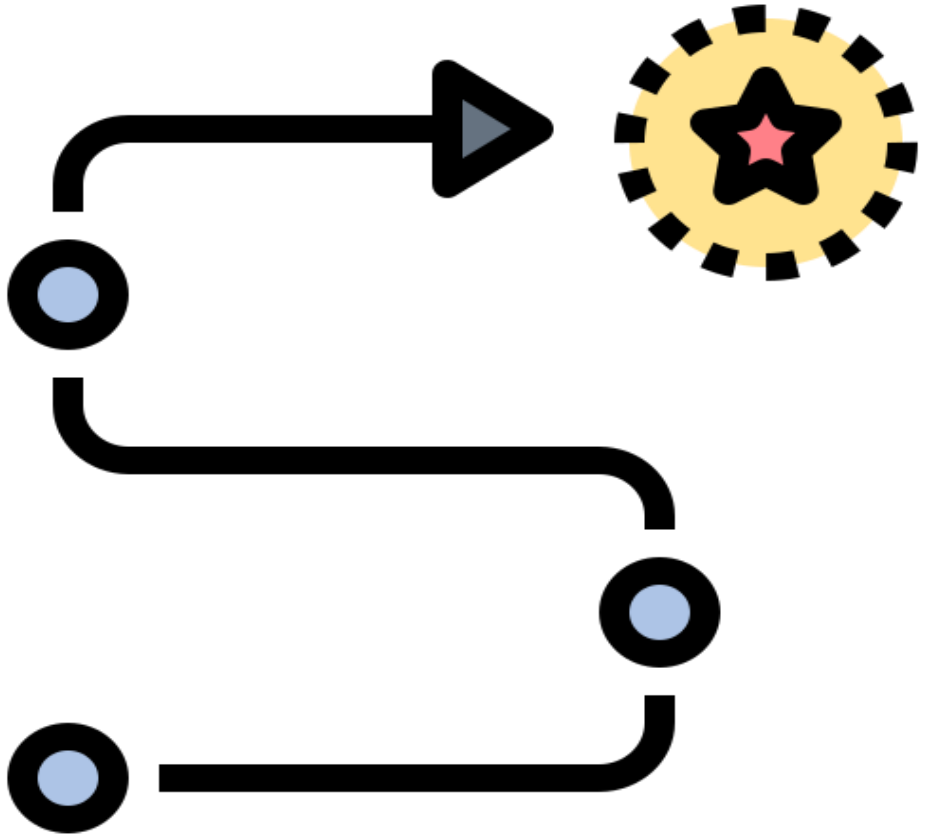
Arabic Language Characteristics (.count)

- The Arabic morphology has some exemptional features: Whereas most languages construct words out of morphemes which are just concatenated, Arabic words are derived using three concepts: root, pattern and form. Generally, each pattern carries a meaning which, when combined with the meaning inherent in the root, gives the goal.

- It doesn't affect the word syntactic category such as verb, noun …etc. Features such as case, gender, number, tense, person, mood and voice are some examples that may be affected by the inflectional morphology

Word	Lemma	Different Interpretations
يعد / yEd/	أعاد />aEAd/	يعِد /yuEid/ (bring back)
	عاد /EAd/	يعُد /yaEud/ (return)
	وعد /waEid/	يعِد /yaEid/ (promise)
	عد /Ead~/	يَعُدّ /yaEud~/ (count)
	أعد />aEd~/	يُعِدّ /yuEid~/ (prepare)

Project Stages



Finding the root directly

Finding the root directly

- The idea is to encode the Arabic letter by our new schema so that helps us simplifies the word so we can get the root.
- we will try to extract possible roots without any dictionary.

First, we will normalize the word
Implement the encoding schema

O: Original letters. These letters are **surely part of the root.** They are:
{ت، ج، ح، خ، د، ذ، ر، ز، ش، ص، ض، ط، ظ، ع، غ، ق}.

P: Prefix letters. These letters **can** be added only in the prefix part. They are:
{ب، ف، س، ل}

S: Suffix letters. These letters **can** be added only in the suffix part. They are:
{ه} only *Haa*

PS: Prefix-Suffix letters. These letters **can** be only added in both sides of the word i.e. in the suffix part or in the prefix part. They are:
{ك، م، ن}

U: Uncertain letters. These letters **can** be added anywhere in the word. They are:
{ت، و، ي، ا، أ}

A: Added letters. These letters are **always** considered additional letters. They are:
{ة} only *Taa Marbuta.*

Finding the root directly (.count)

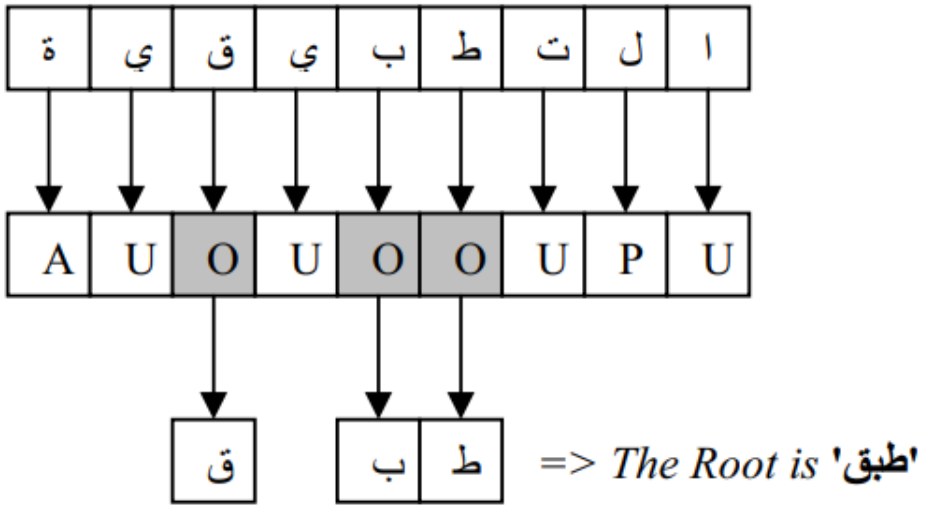
- Then we add position conditions so that help us to find some the prefix and suffix

Table 1- statistics about the position of some letters in the word

The letter	The maximum index in prefixes	The minimum index in suffixes
Baa 'ب'	3	_
Lam 'ل'	5	_
Seen 'س'	4	_
Faa 'ف'	2	_
Haa 'ه'	_	3
Kaf 'ك'	3	*
Noon 'ن'	*	*
Meem 'م'	*	*

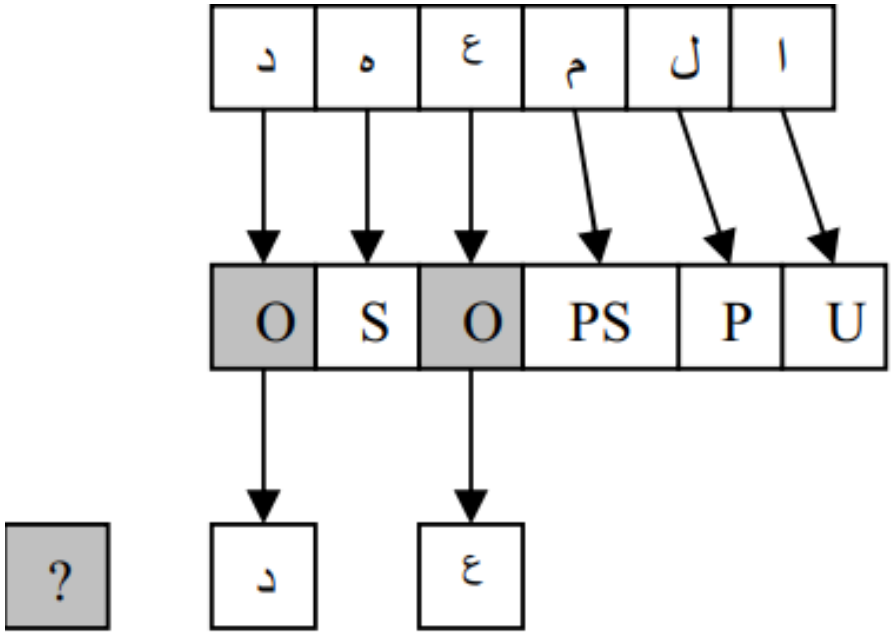
Finding the root directly (.count)

- After implementing all if that the output will look like that:
- if we found 3 “o” then we will stop and that’s our root.



Finding the root directly (.count)

➢ Else like that:



➢Then we implement some Transformation Rules:

R1) Change each 'P' after 'O' to 'O'.

R2) Change each 'S' before 'O' to 'O'.

R3) Change each 'P' after 'S' to 'O', and each 'S' before 'P' to 'O'.

R4) Change each 'PS' before 'P' to 'P'.

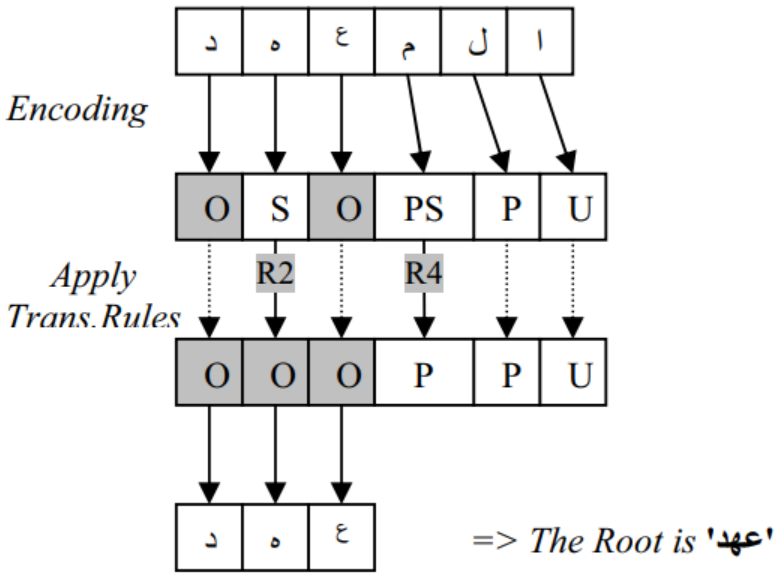
R5) Change each 'PS' before 'O' to 'P'.

R6) Change each 'PS' after 'S' to 'S'.

R7) Change each 'PS' after 'O' to 'S'.

Finding the root directly (.count)

Then the output will look like that :



- After that if we didn't find 3 "o" we have to extract possible roots

- the problem with that we need to make new encoding schema with new rules but because we didn't collect enough rules **we couldn't proceed with this approach**.

FINDING THE POSSIBLE ROOTS

The Current Approach used for the Project

Project Stages

- (**Tokenization**) **-** It starts by splitting the phrase into its words.

ذهب الطالب الى المدرسة

ذهب | الطالب | الى | المدرسة

- Then it checks our pre-made lexicon to see if any of these words is a special word (special words are words that its root is the same as the word itself without changing of its format or pattern)

special_words - Notepad

File Edit Format View Help

الله
من
الي
عن
علي

Word: الى

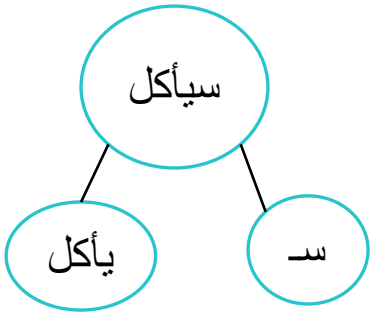
Root: الى

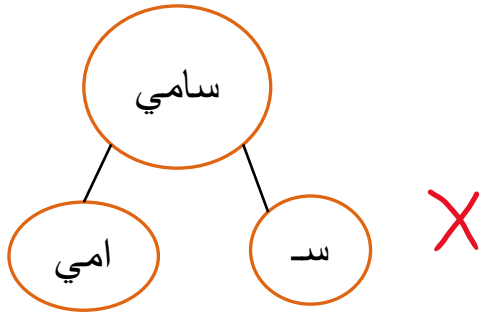
Project Stages (.count)

- But before we continue in the coming steps, we need to normalize the word by removing any diacritics (except the shaddah "~"), numbers,
  and any symbol that isn't an Arabic character ('&' , '"', ',', …).

E.g., كَتَبَ → كتب

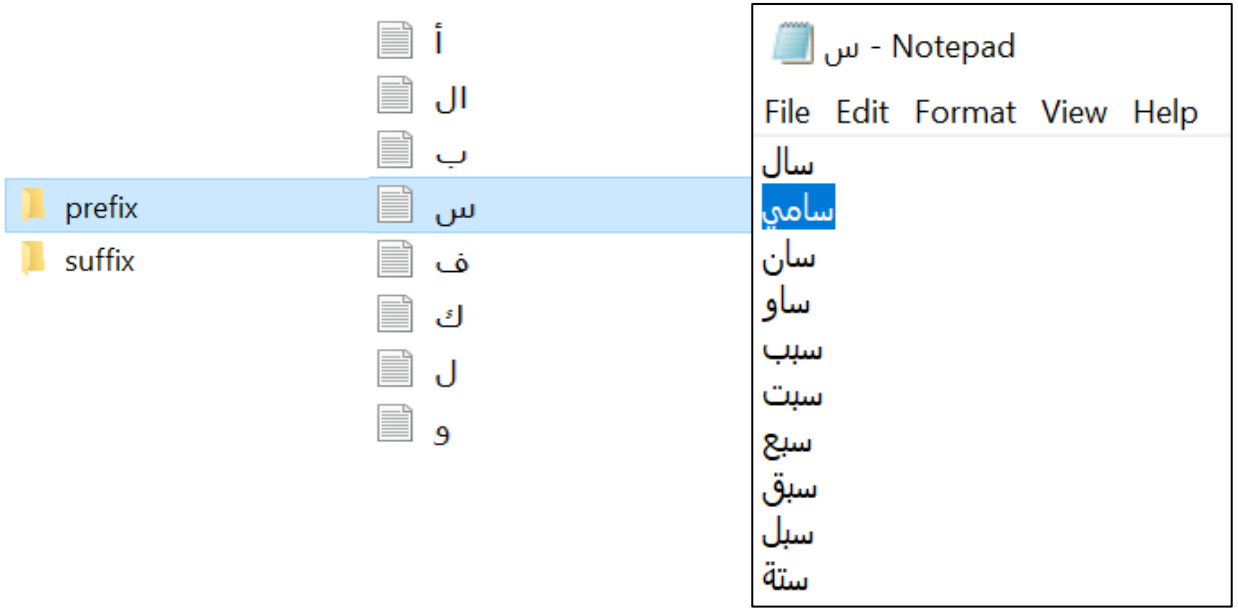
- The next step is to check if the word has any prefixes or suffixes and remove it.
- But before removing any affixes of the word, we should check if these parts are part of the word or not.
- For example,





- This is done by checking the lexicon to see if the specified affix it is part of the word or not

Project Stages (.count)



Word: سامي
Root: سامي

(This process is done for prefixes and suffixes)

- Next step if the affix part is not part of the word.
- Since the word, my have more than one prefix or suffix, then we should be careful when we remove these affixes.
- Then we defined some rules to remove the affixes without remove too many characters that are part of the original words (over stemming), or remove less than needed (under stemming)

RULES

Prefix

ب ك س و ال أ ف ل

---

* س + أ ن ت ي
  سأكتب
  سنكتب
  سيكتب
  ستكتب

س ← الكلمة

بتيجي ملاصقة للكلمة مباشرة

- لو ال "س" جت بعدها الحروف دى، يبقا بنسبة كبيرة ال"س" زيادة
- لو أى حرف تانى، يبقى السين من أصل الكلمة

---

* ال ← الكلمة

ال + ا

- لو الكلمة أولها "ال" يبقا دى من غير rules، هنروح على ال lexicon ونشوف الفايل اللى فيه الكلمات فيه "ال" من أصل الكلمة

---

* أ + أ أأنت؟

أ ← الكلمة

- لو الكلمة بدأت بـ 2 همزة، يبقى أول واحدة همزة زائدة
- لو الكلمة لم تبدأ بـ 2 همزة، يبقى نشوف ال lexicon (لا يوجد قواعد أخرى)

---

* ل + أ ن ت ي
  لأكتب
  لنكتب
  لتكتب
  ليكتب

ل ← الكلمة

- "ل" لو جت قبل الحروف دى يبقى غالباً زيادة

ف ← ل ← الكلمة

"فليجذبوا الناس، نشروا إعلاناتى" كل مكان

الفاء - ممكن يجى بعدها "ل" زائدة، أو تتصل بالكلمة مباشرة

1- وعليه، فهنشوف لو ال "ف" جت بعدها "ل"، يبقا نشوف لو الكلمة [من أول اللام] دى موجودة فى ال lexicon ولا لأ.

2- لو مش موجودة.. نشوف ال "ل" هل هى زائدة ولا لأ (عن طريق الخطوات السابق ذكرها)، لو زائدة ← إذاً الفاء أكيد زائدة غير زائدة (اللام من أصل الكلمة) ← يبقى هنسيب الفاء

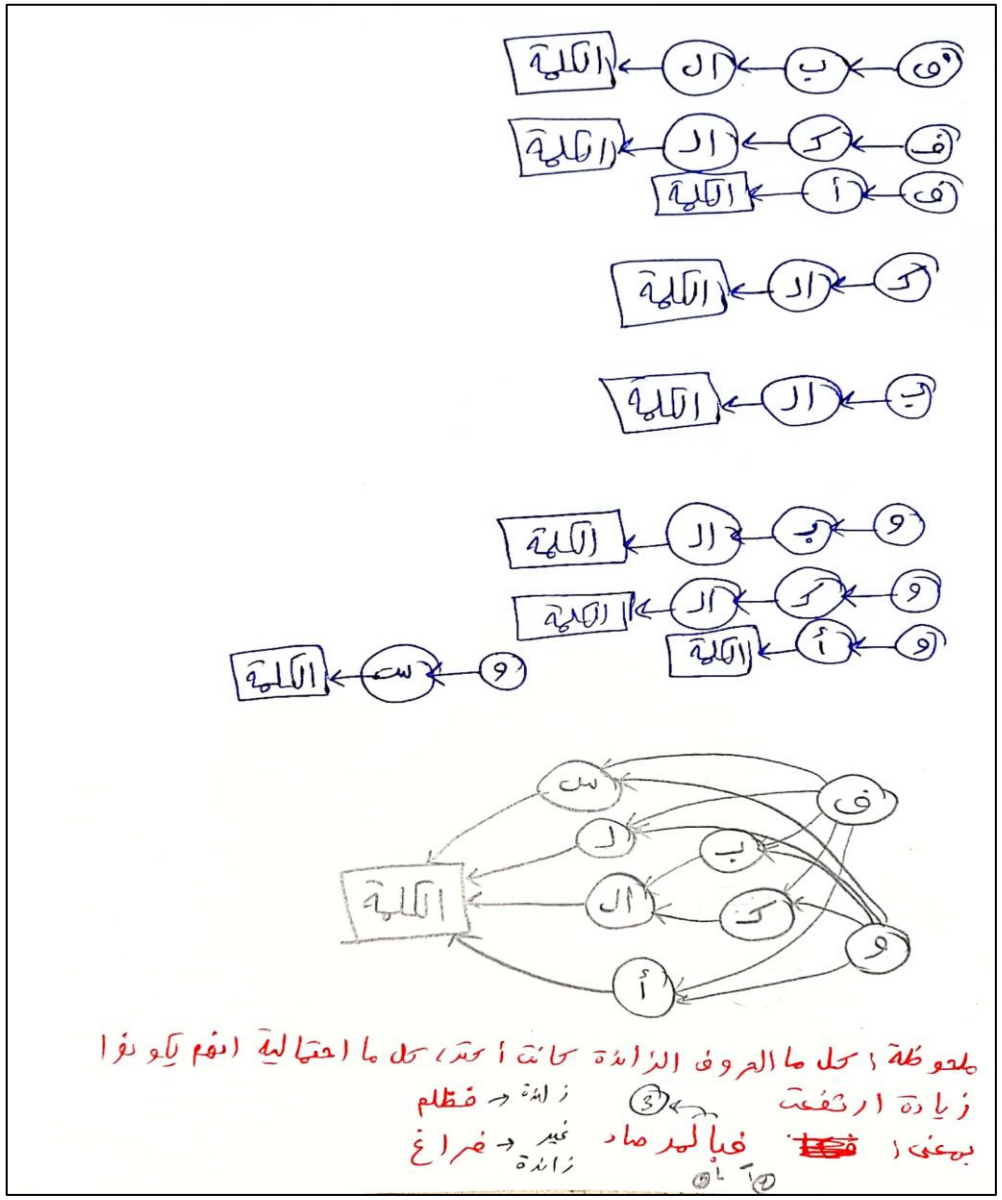
ف ← س ← الكلمة

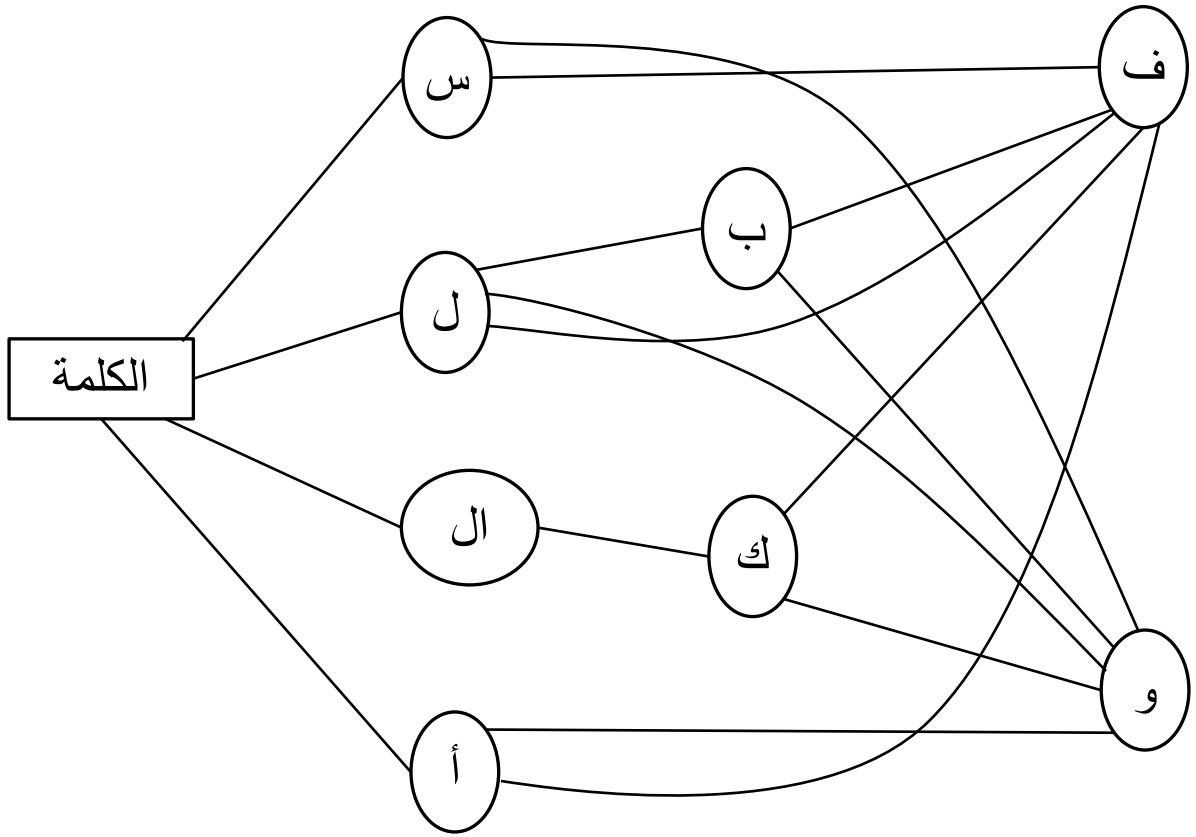
"فستذكرون"

[نفس النظام كما سبق ذكره]

ملحوظة: بعد كل مرحلة بنشوف إذا كانت الكلمة اللى وصلتلها حالياً فى ال lexicon ولا لأ

RULES (.count)





Final diagram for the arrangement of the prefixes with maximum 3 prefixes per word

RULES (.count)

- For the suffixes there are a lot of combinations of the suffixes characters and not have the same arrangements the prefixes, we wrote all possible combinations of the suffixes that we can find.

ا	اء	ء	اءه	ءا	ءه	ءات	ائك	ائكم	ئك
ئكم	ائكما	ائكن	ئكما	كن	ئكن	ائنا	ئنا	ائهم	هم
ئهم	ائهما	ئهما	هما	ما	ائهن	ئهن	هن	ات	اتك
تك	نك	انك	اتكم	تكم	اتكما	تكما	اتكن	تكم	اته
ته	تها	اتها	اتهم	تهم	اتهما	تهما	هما	ما	اتهن
تهن	هن	اك	اكم	اكما	كما	اكن	ان	اه	اها
ها	اهم	هم	هما	اهما	اهن	هن	ة	ت	تا
تاك	اك	تاكم	اكم	تاكما	اكما	تاكن	اكن	تان	تاه
اه	تاها	اها	تاهم	تاهم	اهم	تاهما	اهما	تاهن	اهن
تاي	اي	تك	تك	تكم	تكم	تكما	تكما	تكن	تكن

Pattern Recognition

- The Last stage is the pattern recognition process.
  Where we have a set of patterns and return ones that matches with our word.

- But there is not enough rules, if we choose to return only one answer, it may be in accurate.

- So, we made up for it be returning a list of possible roots for the word.
- With that we increased the accuracy of the program.

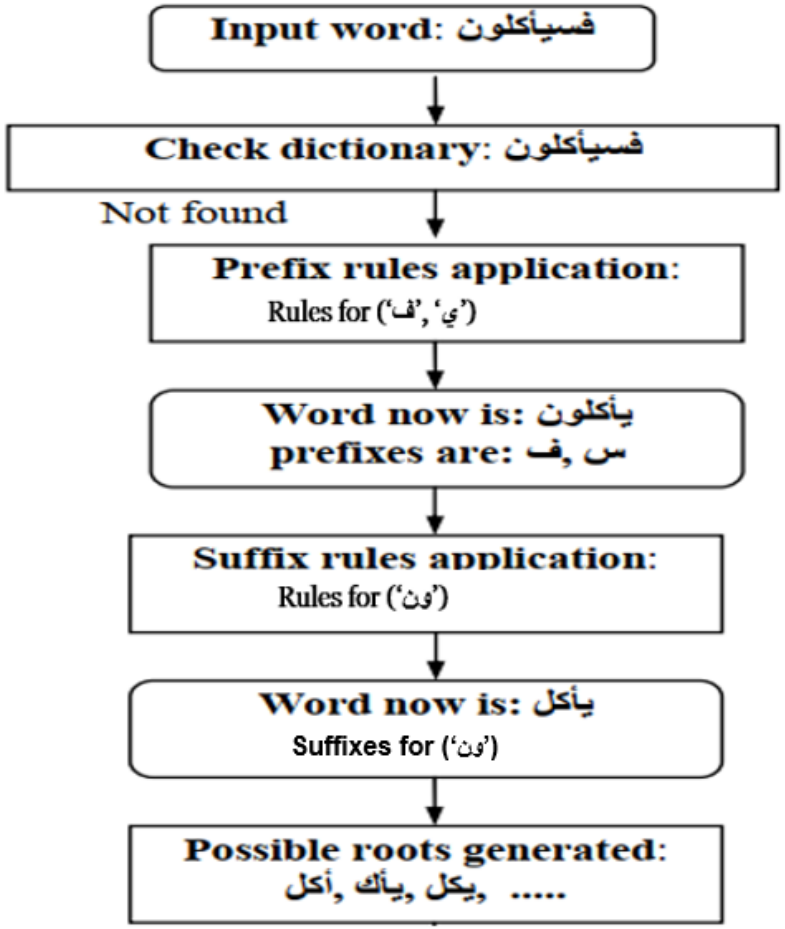
يأكل

Matched patterns: يفعل
Possible roots: أكل

اشترى

Matched patterns: افتعل، افعلل
Possible roots: شرى، شترى

The overflow of the whole process



Additional Features

Finding the type of plural

جمع المذكر السالم:

الجمع :اسم ناب عن ثلاثة فأكثر، بزيادة في آخره، مثل) مدرس، مدرسات (أو تغيير في بنائه مثل) أطفال، وكُتُب، ورؤساء (فالأول يسمى الجمع السالم، والثاني يسمى الجمع المكسر)جمع التكسير .(وينقسم السالم إلى مذكر ومؤنث، وينقسم جمع التكسير إلى جمع قلة وجمع كثرة كما سيأتي الحديث عنه.

تعريف جمع المذكر السالم :هو ما دل على الجمع المذكر من غير تغيير في بناء مفرده مثل (معلمون، ومهندسون، وكاتبون) مفرده سلم من التغيير.

شروط ما يجمع جمع المذكر السالم:يجمع الاسم جمع مذكر سالما إذا كان :

(علماً)لا نجمع طفلاً هذا الجمع لأنه ليس علماً **(مذكراً)**

(لا نجمع سعاد هذا الجمع لأنه ليس مذكراً)

لا نجمع أسداً هذا الجمع لأنه ليس عاقلا

لا نجمع حمزة هذا الجمع لأن في آخره تاء **(خال من التاء في آخره**

(خال من التركيب) :لا نجمع عبد الله أو سيبويه لأنهما علمان مركبان.

Finding the type of plural (.count)

جمع التكسير:

-1**تعريفه :**جمع التكسير) ويُسمّى الجمعَ المُكسر (هو ما دل على الجمع بتغيير في بناء مفرده، ولهذا التغيير ثلاثة أشكال:

- تغيير بالزيادة: طفل أطفال
- تغيير بالنقصان: كتاب كتب
- تغيير الحركات: أسد أُسْد

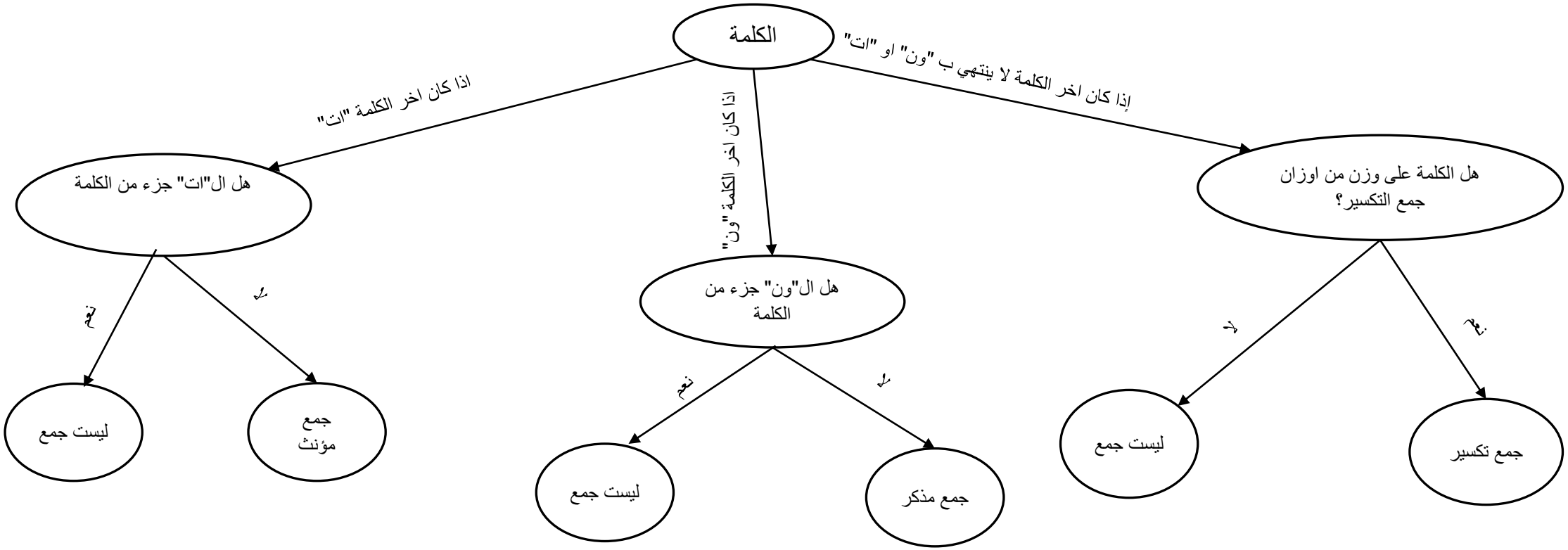
-2**أنواع جمع التكسير:**

أ- جمع القلة :ما دل على العدد القليل، وهو من ثلاثة إلى العشرة

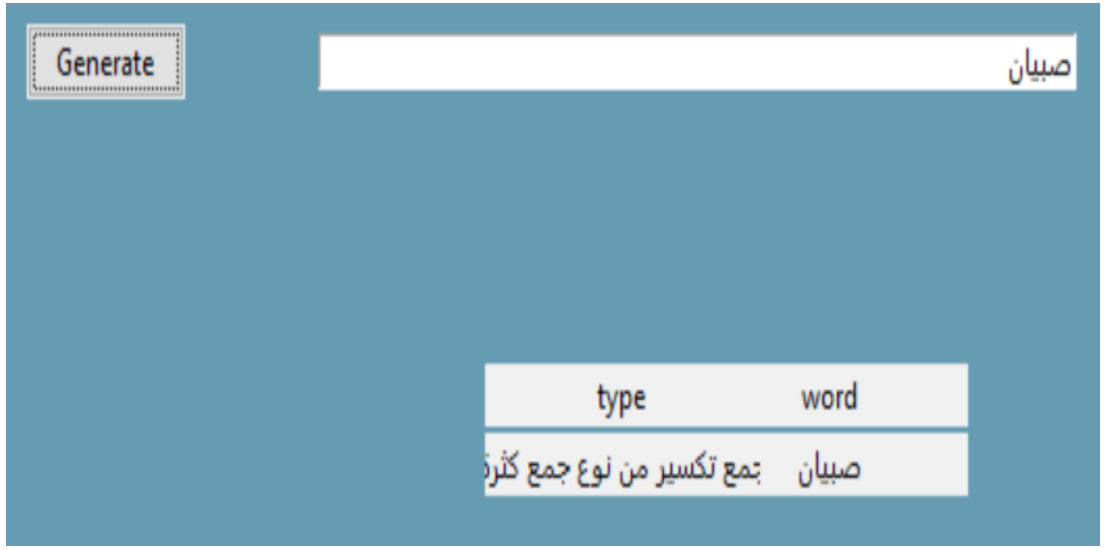
جمع الكثرة :ما تجاوز الثلاثة إلى ما لا نهاية له،

منتهى الجموع:

من جموع الكثرة جمعٌ يُقال له: " منتهى الجموع وهو كل جمع كان بعد ألف تكسيره حرفان، أو ثلاثة أحرف، وسطها ساكن: دراهم، دنانير


الكلمة
اذا كان اخر الكلمة "ات"
اذا كان اخر الكلمة "ون"
إذا كان اخر الكلمة لا ينتهي ب "ون" او "ات"
هل ال"ات" جزء من الكلمة
هل ال"ون" جزء من الكلمة
هل الكلمة على وزن من اوزان جمع التكسير؟
نعم
لا
نعم
لا
نعم
لا
ليست جمع
جمع مؤنث
ليست جمع
جمع مذكر
ليست جمع
جمع تكسير

Finding the type of plural (.count)



Challenges of Arabic Morphological Analyzer

Part of speech Ambiguity

e.g., أكل أحمد كامل الطعام

may be understood as:
- Ahmed ate all the food
- Ahmed Kamel ate the food

Words without Diacritics may bring more than one meaning of the phrase

e.g., علم الولد الدرس

→ علَّم==Teach
→ علِم== Knew

Ignored Tasks

Machine Learning Approach

- We tried to use the first approach of the morphological analyzer approaches (Data-Driven Approach) and train a machine learning model.

- So, we collected data from a data source to train the data on it.

Machine learning approach (.count)

←T→	id	vocalized	unvocalized	wordtype	root	normalized	stamped	original	mankous	feminable	number
Edit Copy Delete	4	وَاحِدٌ	واحد	اسم فاعل	وحد	واحد	حد			Ta	مفرد
Edit Copy Delete	5	وُحْدَانٌ	وحدان	اسم فاعل		وحدان	حدن				جمع تكسير
Edit Copy Delete	6	بَارٌّ	بار	اسم فاعل	برر	بار	بر			Ta	مفرد
Edit Copy Delete	7	أَبْرَارٌ	أبرار	اسم فاعل		ءبرار	برر				جمع تكسير
Edit Copy Delete	8	بَرَرَةٌ	بررة	اسم فاعل		بررة	بررة				جمع تكسير
Edit Copy Delete	9	عَامِلٌ	عامل	اسم فاعل	عمل	عامل	عمل			Ta	مفرد

Showing rows 0 - 24 (23019 total, Query took 0.0004 seconds.)

But since the dataset is not big enough, the model we trained returns inaccurate answers and has very low accuracy.
So, we ignored this approach.

Name Entity Recognition

- Named entity recognition (NER) – sometimes referred to as entity chunking, extraction, or identification – is the task of identifying and categorizing key information (entities) in text. An entity can be any word or series of words that consistently refers to the same thing. Every detected entity is classified into a predetermined category.

e.g., **ذهب عادل الى مكة**
ذهب ~ O
عادل ~ Person
الى - O
مكة - Place

Name entity recognition (.count)

- Here we searched for a pre-defined NER model and try to link it to our program.
- We use the dataset from AQMAR Arabic Wikipedia Named Entity Corpus

```
x = ""
lstm_predict(x)
```

[22] ✓ 0.1s

```
كان : O
أستاذ : O
أحمد : B-PER
عادل : O
يذهب : O
يوميا : O
الى : O
عمله : O
في : O
مدينة : O
الأسكندرية : B-LOC
```

- But had an error that we couldn't handle, and the model wasn't good enough, so, we ignored this feature.

References

[1] Riyad Alshalabi (2005). Pattern Based Stemmer of Finding Arabic Roots, Information Technology Journal 4 (1): 38-43.

[2] TENGKU MOHD T. SEMBOK, & BELAL MUSTAFA ABU ATA &ZAINAB ABU BAKAR (2011). A Rule-Based Arabic Stemming Algorithm, Conference Paper. Research Gate.

[3] Mohamed Boudchiche & Azzeddine Mazroui & Mohamed Ould Abdallahi Ould Bebah & Abdelhak Lakhouaja, Abderrahim Boudlal (2017). AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer, Journal of King Saud University – Computer and Information Sciences.

[4] Riad Sonbol & Nada Ghneim & Mohammed Said Desouki. Arabic Morphological Analysis: A New Aprroach, Informatics Department, HIAST Damascus, SYRIA.

[5] Ameerah Alothman1 & AbdulMalik Alsalman (2020). Arabic Morphological Analysis Techniques, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 2.

[6] الغلاييني، جامع الدروس العربية

[7] سعيد الأفغاني، الموجز في قواعد اللغة العربية

Thanks