Faculty of Computer and Information

Cairo University

Arabic TTS

Team Names:

Samir Mohamed

Ali Abd Elrahman

Mai Ahmed

Amira AbdElNaby

Dr. Hanaa Bayomi

TA. Amr Magdy

Project Development

2.1 Work plan for the project that should specify each task

(In details)

Task name	Description
1- Collect Arabic data with Tashkeel.	We collect huge amount of Arabic data with all tashkeel.
2- Test Mbrola tools	Mbrola tool is a tool the take Arabic data without tashkeel and pronounce it, we test each char with all possible tashkeels and it accuracy is very low.
3- Split each file to files	Split each large file o smaller size to be able to use it.
4- Split each file to sentence.	We go for each file and split each file to sentence.
5- Do Modawwana code	At first we have 2 tables (primary and secondary) the code : • Remove the symbols • Remove the tashkeel • Add the word in primary if it was not there and update its counter. • Then add it with tashkeel in secondary and update its counter.
1- Improve the Modawwana code .	Algorithm: • We use indexing and hash Map for each character in Arabic. • There are 2 tables for each character, primary and secondary. • The primary for word that begins with that character without tashkeel. • The secondary for word that begins with that character

	with tashkeel. • In primary table, there are 3 records: id, word and counter for that word in all our data. • In secondary table, there are 4 records: id and word, counter for that word in all our data and foreign key for this word in primary table. • First, we will iterate for each character in Arabic in all our data files. • We load data for that character from database and manipulate it in hash Map (insert and update). • Then, get all words from selected file and loop in them. • For each word, keep that word and make new one without tashkeel. If first character in word matches with selected character, there are 3 scenarios : 2- If word is not in primary table, it will be added in primary and secondary table. Note: there are 3 words for highlighting the word be selected: • "New" for highlighting that word is new in database and call insert query. • "Update" for highlighting to update counter of this word and call update query. • "Old" for highlighting to do nothing for this word. 3- If word in primary table and not in secondary table, it will be added in secondary table only and increase word's counter in primary table. 4- If word in primary and secondary

	tables, counter of that word will be updated in both tables.
5- Run the Files in the Modawwana code .	We go to each file and run it in the Modawwana code to insert it to the data base.
6- Do frequency code.	We made a code to insert in the database how many times each word is repeated with another in a specified window.
7- Run the files in the frequency code.	We go to each file and run it in the frequency code to insert it to the data base.
8- Calculate the number of occurrence of two words with each other.	Getting the counter of 2 words which they repeated with each other in specific window size in Python code.
9- Calculate accuracy.	We made a code to run the Naive Bayes code using testing file and then remove the tashkeel from it and compare the result with the original one and compute the accuracy.
10- Choice one of (Recurrent neural network , Apply Long short term model , CRF++)	We will substitute the Naïve Bayes algorithm with one of the listed algorithms based on what will give higher accuracy.
11- Replace Absalom with N-gram.	We will use N-gram.
12- Integrate it with E-speak.	We will integrate our code with E-speak (an open source code to speak Arabic).
13- Increase the amount of training data	We will increase the number of files used in training phase.
14- Using development data	We will use dev data to enhance the code and change the features and observe the result trying not to produce fake accuracy while using test data.
15- Represent accuracy by graph in excel sheet.	We change the window and calculate the accuracy for each one and represent it by graph in excel sheet.

2.2 Current state:

2.2.1 Description of current state:

We make a primarily design that take a sentences without tashkeel and add the right tashkeel to it.

Accomplished Task name	Time
1- Collect Arabic data with Tashkeel.	1 Month
2- Test Mbrola tools	1 Week
3- Split each file to sentence.	3 Weeks
4- Remove tashkeel and samples from the sentences.	3 Days
5- Do Modawwana code .	2 weeks
6- Improve the Modawwana code .	1 Weeks
7- Run the files in the Modawwana code.	2weeks
8- Do frequency code.	1 Week
9- Run the files in frequency code.	2 weeks
10- Calculate the number of occurrence of two words with each other.	3 Weeks
11- Python code	1 week
12- Python database	1 week

2.3 The rest of tasks with approximately taken time

Unaccomplished Task name	Time (approximately)
1- Calculate accuracy.	4 days
2- Choice one of (Recurrent neural network , Apply Long short term model , CRF++)	1 month
3- Replace Absalom with N-gram.	1 week
4- Integrate it with E-speak.	2 Weeks
5- Increase the amount of training data	1 week
6- Using development data	4 days
7- Represent accuracy by graph in excel sheet.	7 days