الجمهورية الجزائرية الديمقراطية الشعبية

Ministry of Higher Education and Scientific Research

DR. YAHIA FARES UNIVERSITY OF MEDEA

Faculty of Science

Department of Mathematics and Computer Science



Graduation Project

Submitted for graduation in

Computer science master

Specialty: Systems Engineering and Web Technology

TITEL

Application of Word2Vec Algorithms to Improve Queries of Arabic Question Answering System.

Prepared and Presented by

- ✓ HACHEMI Mohamed Ramzi
- ✓ IRKI Amira

Proposed and directed by

- ✓ MAZARI Ahmed Cherif

Jury Members

- ✓ Ouled Baba Ali Kawther
- ✓ Abdoun Nabil

Academic year : 2017 – 2018

Abstract

Arabic is a rich language, but there have not been many opportunities to use it in the field of scientific research. For this we suggested a way to implement the Word2Vec algorithm to improve Arabic Questions Answering System, it was tested on a set of questions written in Arabic.

Word2Vec created by Google in 2013, Word2Vec use as an entry a large or small text and then produce a vector space usually of several hundred dimensions.

Word2vec has been used in several domains, from these domains Question answering system with different language. But we noted that these works are not very applied to our mother language "Arabic language".

We suggested a way for implement Word2Vec algorithm to improve Arabic Question answering system, it was tested on a set of questions written in Arabic.

Our method is to improve questions in Arabic language, then apply the Word2Vec algorithm with its models, on this questions after doing a preprocessing to extract similar words in context or meaning, then try to find the answers to the reformulated questions via Google API. Only the relevant answers are kept for each question.

Other times, we take the initial questions without preprocessing (which means in the original state) and try to find the answers also via Google API, and we keep just the relevant answers.

From the results we get after doing the two parts, we compare between the two results (initial and reformulated), to see if there is an improvement or not according to the percentage.

Keywords: Arabic language, Word2Vec, Question Answering System, Similarity, Deep Learning.

ملخص

اللغة العربية هي لغة غنية ، ولكن لم تكن هناك الكثير من الفرص لاستخدامها في مجال البحث العلمي. لهذا اقترحنا طريقة لتطبيق خوارزمية Word2Vec لتحسين نظام الإجابة على الأسئلة العربية ، وتم اختبارها على مجموعة من الأسئلة المكتوبة باللغة العربية.

Word2Vec تم إنشاؤه من قبل Google في عام 2013، يستخدم Word2Vec كمدخل نص صغير أو كبير ثم يولد الفضاء ناقلات عادة عدة مئات من أبعاد.

تم استخدام Word2vec في العديد من المجالات ، من هذه المناطق: نظام الإجابة على الأسئلة وبلغات مختلفة. لكننا لاحظنا أن هذه الأعمال لا تطبق بشكل كبير على لغتنا الأم "اللغة العربية".

اقترحنا طريقة لتطبيق خوارزمية Word2Vec لتحسين نظام الإجابة على الأسئلة ، ومن ثم اختبارها على مجموعة من الأسئلة المكتوبة باللغة العربية.

تتمثل طريقتنا في تحسين الأسئلة باللغة العربية ، ثم تطبيق خوارزمية Word2Vec بنماذجه ، بعد إجراء معالجة مسبقة لاسترداد كلمات مماثلة في سياقها أو معناها ، ثم محاولة العثور على إجابات للأسئلة المُعاد صياغتها عبر Google API ثم يتم الاحتفاظ فقط الإجابات ذات الصلة لكل سؤال.

في أحيان أخرى ، نأخذ الأسئلة الأولية دون معالجة مسبقة (وهذا يعني في الحالة الأصلية) ونحاول العثور على الإجابات أيضًا عبرإستعمال Google API ، ونحتفظ فقط بالإجابات ذات الصلة.

من النتائج التي حصلنا عليها ، بعد إجراء كلا الجزأين ، نقارن بين النتيجتين (الأولية والمعدلة) ، لمعرفة ما إذا كان هناك تحسن أم لا وفقا للنسبة المئوية.

الكلمات الرئيسية: اللغة العربية ، Word2Vec ، الإجابة على سؤال النظام ، التشابه ، التعلم العميق.

Résumé

L'arabe est une langue riche, mais il n'y a pas eu beaucoup d'occasions de l'utiliser dans le domaine de la recherche scientifique. Pour cela nous avons suggéré un moyen de mettre en œuvre l'algorithme Word2Vec pour améliorer le système de réponse aux questions arabes, il a été testé sur un ensemble de questions écrites en arabe.

Word2Vec créé par Google en 2013, Word2Vec utilise comme entrée un petit ou grand texte et produit ensuite un espace vectoriel généralement de plusieurs centaines de dimensions.

Word2vec a été utilisé dans plusieurs domaines, à partir de ces domaines Système de réponse aux questions avec des différents langages. Mais nous avons noté que ces travaux ne sont pas très appliqués à notre langage maternelle "langue arabe".

Notre méthode est d'améliorer les questions en arabe, puis d'appliquer l'algorithme Word2Vec avec ses modèles, après avoir effectué un prétraitement pour extraire des mots similaires dans leur contexte ou leur signification, puis essayer de trouver les réponses aux questions reformulées via Google API. Seules les réponses pertinentes sont conservées pour chaque question.

D'autres fois, nous prenons les questions initiales sans prétraitement (ce qui signifie dans l'état original) et essayons de trouver les réponses aussi via Google API, et nous gardons juste les réponses pertinentes.

A partir des résultats que nous avons obtenons, après avoir fait les deux parties, nous comparons entre les deux résultats (initial et reformulé), pour voir s'il y a une amélioration ou non selon le pourcentage.

Mots-clés: Langue arabe, Word2Vec, Question Answering System, Similarité, Apprentissage profond.

Thanks

At the beginning, we thank God Almighty who provided us with the strength necessary to accomplish this work successfully.

We would like to express our deep gratitude to our adviser, Mr. Mazari, for his constant guidance, challenged discussions, tips, enthusiasm and knowledge. It stimulated us to think more deeply about our work. We also made a great effort to build up the structure and improve every detail of our work. We are grateful to him for working with us. We learned a lot, it was an honor.

Also, we would like to thank طه زروقي, Mr. Abu Bakr Soliman, Mr. Benajiba, Mr. Mazari Reda, Mr. Sayh Omar, and Mr. Zerrouki for their help to complete our work, because they give all the information needed

We express our deep appreciation to the members of the jury for their interest in accepting and reviewing our work.

Our eternal gratitude to our parents who encouraged us to be the best we can be and we have high expectations and constant prayer for our success.

Also, we are grateful to our entire family and friends for their encouragement and support.

We would also like to thank all the professors of Medea who helped us achieve what we are now.

We also sincerely thank the Information Department for facilitating the means to accomplish this work.

Finally, we would like to thank all those who participated in or near this work.

"We hope you like it our memory."

Médéa, June 2018

Amira, Ramzi



Dedication

*"Praise **God,** the one and only".*

That this modest work, a little towards others bigger,

But he will always remain modest before your mercy.

I dedicated this modest work

*To my dear **mother** and **father**,*

Who always supported during my life, my studies, and encourage moving forward.

***To my sisters** (Kenza, Nadjet).*

*To the **little baby** of family (Mohammed Haythem).*

***To my partner** (Ramzi).*

*To all my **close friends** (Yasmine, Naziha, Souad, Yacine, hadda, Wafa, malek, Wasna).*

*To the whole Irki **family**.*

*To all the **promotion** of 2nd year master's degree in Computer Science (2017/2018).*

*To **everyone** who knows me.*





«I.mira»

Dedication

*"Praise **God,** the one and only".*

That this modest work, a little towards others bigger,

But he will always remain modest before your mercy.

I dedicated this modest work

*To my dear **mother** and **father**,*

Who always supported during my life, my studies, and encourage moving forward.

***To my Sister** "Mother of abdel ilah" and their childs: "abdel illah" and "Rahil"with **my Brothers** "Hamza"**and**"Soufiane".*

To all Hachemi (hamouli) members

*To all my **close friends** (Yassine,Ghanou,Krimou,Yasser,Haroune,Mohamed,abd Malek,Imad,Kader,Rachid,Otti,Mahdi,Ali,Lotfi,Rekai).*

To all my friends from others cities

***To my Futur wife** "Sondrine".*

***To my partner** (Amira).*

*To all the **promotion** of 2nd year master's degree in Computer Science (2017/2018).*

*To **everyone** who knows me.*



«H.Rmimez»

Table of content

Chapter I: State Of The Art……..……..1
1. Introduction: 2
2. Deep Learning: 2
2.1. Deep Learning Definition: 2
2.2. Difference between Neural Network and Deep Learning Neural Network: 2
2.3. What Is the Difference between Deep Learning and Machine Learning?: 3
2.4. Deep Learning type: 4
2.4.1. Deep Learning for Unsupervised Learning: 4
2.4.2. Convolutional Neural Networks: 4
2.4.3. Restricted Boltzmann Machine: 4
2.4.4. Recursive Neural Network: 4
2.4.5. Recurrent neural network (RNN): 4
2.4.6. Autoencoders: 5
2.4.7. Deep Belief Networks (DBN): 5
2.5. Deep learning application and examples: 5
3. Question Answering System: 5
3.1. Question answering (QA) System Definition: 5
3.2. Question answering (QA) system History: 5
3.2.1. BASEBALL [Green et al., 1963]: 5
3.2.2. LUNAR [Woods, 1977]: 5
3.2.3. STUDENT: 6
3.2.4. PHLIQA: 6
3.2.5. UC (Unix Consultant): 6
3.2.6. LILOG: 6
3.2.7. Open-domain QA: START [Katz, 1997]: 6
3.2.8. WolframAlpha 2009: 6
3.2.9. IBM Watson [1] wins at Jeopardy 2011: 6
3.3. Question answering (QA) system Types: 7
3.3.1. Open-domain: domain independent QA systems can answer any query from any corpus: 7
3.3.2. Closed-domain: domain specific QA systems are limited to specific domains: 7
3.4. Difference between QAS and search engine: 7
3.5. Question answering (QA) system Applications: 7
3.5.1. Personal assistants: 7
3.5.2. Access to Encyclopedic Knowledge: 8
3.5.3. Smart Home and Entertainment: 8
3.5.4. Search: 8
3.6. Question answering (QA) system Architecture: 8
3.6.1. Questions Categories: 9
3.6.2. Natural Language Processing: 9

3.6.2.1. Question Analysis: 9
3.6.2.1.1. Question Analysis: Named Entity Recognition: 9
3.6.2.1.2. Question Analysis: Pattern Learning: 10
3.6.2.1.3. Question Analysis: Syntactic Parsing: 10
3.6.2.1.4. Question Analysis: Semantic Role Labeling: 10
3.6.2.2. Question Classification: 10
3.6.3. Information Retrieval: 11
3.6.3.1. Query Construction: 11
3.6.3.2. Document Retrieval: 12
3.6.3.3. Sentence Retrieval: 12
3.6.3.4. Sentence Annotation (Natural Language Processing): 12
3.6.4. Information Extraction: 13
3.6.4.1. Type of Answer: 13
3.6.4.2. Answer Extraction: 13
3.6.4.3. Answer Validation: 14
4. Conclusion: 14
Chapter II:Word2Vec……...15
1. Introduction: 16
2. Word embeddings and word2vec: 16
2.1. Vector Space Model Representation: 16
2.1.1. Representation: 16
2.1.2. Notion of Relatedness: 17
2.2. Word Embedding Definition: 17
2.2.1. Word Embeddings History: 18
2.2.2. Why Word Embedding Models Are Used? : 18
2.2.3. Word Embedding Models: 18
2.3. Word2Vec Definition: 18
2.3.1. Continuous Bag of Words (CBOW) Model: 18
2.3.2. Skip Gram Model: 19
2.4. Word2Vec Applications: 20
2.4.1. Word Similarity: 20
2.4.2. Machine Translation: 20
2.4.3. Part-of-Speech and Named Entity Recognition: 20
2.4.4. Relation Extraction: 20
2.4.5. Sentiment Analysis: 21
2.4.6. Co-reference Resolution: 21
2.4.7. Clustering: 21
2.4.8. Semantic Analysis of Documents: 21
2.5. Word2vec use in our world: 22
2.6. Word2Vec structure: 22
2.6.1. Define the entry and the exit of Word2Vec (Input/Output): 22
2.6.2. Tokenization: 22

2.6.3. Turning Sentences into Vectors and Count Number of existing word in vectors: 23
2.6.4. The hidden layer: 23
2.6.4.1. Choosing from Word2Vec Architecture (CBOW or Skip Gram): 23
2.6.4.2. Training Neural Network: 24
2.6.5. Training algorithm: 25
2.6.5.1. Hierarchical Softmax: 25
2.6.5.2. Negative sampling: 25
2.6.5.2.1. How the negative samples are chosen?: 26
2.6. 6.Final Output of word2vec: 26
2.6.7. Drawing the graph: 26
2.6.8. Word Similarity: 27
3. Conclusion: 28
Chapter III: Deep Learning For Arabic NLP……..…29
1. Introduction: 30
2. Natural Language Processing (NLP): 30
2.1. Natural Language Processing (NLP) Definition: 30
2.2. Natural Language Processing (NLP) Steps: 30
2.2.1. Lexical Analysis:. 30
2.2.2. Syntactic Analysis (Parsing): 30
2.2.3. Semantic Analysis: 31
2.2.4. Discourse Integration:. 31
2.2.5. Pragmatic Analysis: 31
2.3. Natural Language Processing (NLP) Application: 31
2.3.1. Machine Learning: 31
2.3.2. Human Computer Interaction: 32
2.3.3. Information Extraction: 32
2.3.4. Summarization: 32
2.3.5. Reduce words to their root (or stem): 32
2.4. Natural Language Processing systems (NLP) Examples in artificial intelligence: 32
2.4.1. Communication: 32
2.4.2. Faster diagnosis: 33
2.4.3. Customer Review: 33
2.4.4. Virtual digital assistants: 33
3. Arabic Natural Language Processing (ANLP): 33
3.1. Arabic Language Definition: 33
3.2. Arabic Language Types: 34
3.2.1. Classical Arabic CA: 34
3.2.2. Modern Standard Arabic MSA:. 34
3.2.3. Dialect Arabic DA: 34
3.3. Arabic Language Script: 34
4. Works on Deep learning for Arabic NLP: 36
4.1. Optical Character Recognition (OCR): 36

4.1.1. Handwritten Text: 37
4.1.2. Printed Text: 37
4.1.3. Video-Overlaid Text: 37
4.1.4. Natural Scene Text: 38
4.2. Caption Generation: 38
4.3. Automatic Speech Recognition (ASR): 38
4.4. Language Modeling: 39
4.5. Automatic Machine Translation (ATM): 39
4.6. Text Categorization (TC): 40
4.7. Sentiment Analysis (SA): 40
4.8. Question Answering (QA): 40
4.9. Distributed Word Representations: 41
5. Related work: 47
6. Conclusion: 49
Chapter IV: Proposed Approach……..…….……..50
1.Introduction: 51
2. Proposed approach: 51
2.1.Dataset: 52
2.1.1. Dataset Definition: 52
2.2. Questions: 52
2.2.1. Question Definition: 52
2.2.2. Question Type: 52
2.2.3. Question part: 53
2.3. Unique encoding of texts: 53
2.4. Questions reformulation process: 53
2.4.1. Normalization: 54
2.4.2. Skip words between quotation marks: 55
2.4.3. Remove "Stop words": 55
2.4.4. Skip Named entities: 55
2.4.4.1. Named entity Definition: 55
2.4.5. Put Named Entity between quotation marks: 56
2.4.6. Transformation number to string (Int to Str): 56
2.4.7. Transformation string to number (Str to Int): 56
2.4.8. Extract similar words: 56
2.5. Two models of Word2Vec: 57
2.5.1. Continuous bag-of-words (CBOW): 57
2.5.2. Skip gram: 57
2.6. Similarity: 57
2.4.9. Stemming (similar words): 58
2.4.10. Determination the threshold of similarity: 58
2.4.11. Use of "+"sign or " " (Space) sign: 58
2.7. Get answers from IRS: 58

2.8. Testing the results: 59
2.8.1. Test1 (relevant/irrelevant): 59
2.8.2. Test2 (improvement/not improvement): 60
3. Conclusion: 61
Chapter V: Implementation And Evaluation……..…….62
1. Introduction: 63
2. Development environment: 63
2.1. Python: 63
2.1.1. Python Features: 63
2.1.1.1. Open source: 63
2.1.1.2. Easy-to-learn: 63
2.1.1.3. High-level Language: 63
2.1.1.4. Portable: 63
2.1.1.5. Object-Oriented: 63
2.1.1.6. Python is Interactive: 63
2.1.1.7. Interpreted: 64
2.1.1.8. Extendable: 64
2.1.1.9. Libraries: 64
2.1.1.10. Supports: 64
2.2. Python module used: 64
2.2.1. Gensim Module: 64
2.2.2. Re Module: 64
2.2.3. Unicodedata Module: 64
2.2.4. Pyarabic Module: 65
2.2.4.1. Features: 65
2.2.5. Python string Module: 65
2.2.6. Google APIs Module: 65
2.2.7. Matplotlib Module: 66
2.2.8. Tahsphyne Module: 66
2.2.8.1. Features: 66
2.2.9. PANDA Module: 66
2.2.10. Numpy Module: 66
2.2.11. Scikit-learn: 67
3. Explanation of our approach: 67
3.1. Benajiba Dataset: 67
3.2. Benajiba questions: 67
3.2.1. Criticizing of Benajiba Dataset: 68
3.2.2. Question part: 68
3.3. Use of Farasa Dataset: 69
3.4. Unique encoding of texts: 70
3.5. Unicodedata: 70
3.6. Question reformulation process: 70

3.6.1. Normalization: 70
3.6.2. Skip words between quotation marks: 71
3.6.3. Remove "Stop words": 71
3.6.4. Skip Named entities: 72
3.6.5. Put Named Entities between quotation marks: 72
3.6.6. Transformation number to string (Int to Str): 72
3.6.7. Transformation string to number (Str to Int): 72
3.6.8. Extract Similar Words to the rest words. 72
3.7. Models of Word2Vec: 73
3.8. Similarity: 73
3.6.9. Stemming for similar words: 74
3.9. Get answers of questions: 74
3.6.10. Determination the threshold of similarity: 74
3.6.11. Use of "+"sign or " " (Space) sign: 75
3.10. Test results (relevant/irrelevant): 75
3.11. Test results (improvement / no improvement): 75
4. Work presentation: 76
4.1. Flow: 76
4.1.1. Test 01: 76
4.1.1.1. Results: 76
4.1.1.2. Discussion: 78
4.1.1.3. Suggestions: 79
4.1.2. Test 02: 79
4.1.2.1. Results: 79
4.1.2.1.1. Results of CBOW model: 80
4.1.2.2. Discussion of CBOW model: 80
4.1.2.1.2. Results of Skip-Gram model: 81
4.1.2.3. Discussion of Skip-Gram: 82
4.1.2.4. Suggestions: 83
4.1.3. Test 03: 83
4.1.3.1. Results: 83
4.1.3.1.1. Results of CBOW model: 84
4.1.3.2. Discussion of CBOW model: 84
4.1.3.1.2. Results of Skip-Gram model: 85
4.1.3.3. Discussion of Skip-Gram model: 85
4.1.3.4. Suggestions: 86
4.1.4. Test 04: 86
4.1.4.1. Results: 86
4.1.4.1.1. Results of CBOW model: 88
4.1.4.2. Discussion of CBOW model: 88
4.1.4.1.2. Results of Skip-Gram model: 88
4.1.4.3. Discussion of Skip-Gram: 88

4.1.4.4. Suggestions: 89
4.1.5. Test 05: 89
4.1.5.1. Results: 89
4.1.5.1.1. Results of CBOW model: 91
4.1.5.2. Discussion of CBOW model: 91
4.1.5.1.2. Results of Skip-Gram model: 91
4.1.5.3. Discussion of Skip-Gram model: 91
4.1.5.1.3. Results of comparison: 92
4.1.5.3. Discussion results of comparison: 94
4.1.5.4. Suggestions: 94
4.1.6. Test 06: 95
4.1.6.1. Results: 95
4.1.6.2. Discussion: 95
4.1.6.3. Queries transformation for CBOW model: 95
4.1.6.4. Queries transformation for Skip-Gram: 95
4.1.6.5. Suggestions: 96
4.1.7. Test 07: 96
4.1.7.1. Results: 96
4.1.7.2. Discussion: 96
4.1.7.3. Queries transformation for CBOW model: 97
4.1.7.4. Queries transformation for Skip-Gram: 97
4.1.7.5. Suggestions: 97
4. Evaluation of results: 97
5. Presentation of the application: 98
5.1. HTML (HyperText Markup Language): 98
5.2. CSS(Cascading Style Sheet): 99
5.3. JavaScript: 99
5.4. Django: 99
6. Conclusion: 100
General conclusion and perspectives 102
References 103
Web references 110
Annex 114

List of figures

Figure 1: Architecture of Simple Neural Network and Deep Learning Neural Network. [3].3
Figure 2: Difference between Machine Learning and Deep Learning. [2] 4
Figure 3: WolframAlpha Question Answering (QA) System. [Lehmann, 2016] 6
Figure 4: Competition between Jeopardy against human experts. [Lehmann, 2016] 6
Figure 5: Type of Personal assistants (Apple Siri, Google Now, and Window Cortana). [Lehmann, 2016] 7
Figure 6: Smart Home Control. [Lehmann, 2016] 8
Figure 7: Question answering (QA) system Architecture. [Neves, 2014] 9
Figure 8: Syntactic Parsing model. [Neves, 2014] 10
Figure 9: Type of Classification. [Neves, 2014] 11
Figure 10: Different wordings possible, similar syntactic structure. [Neves, 2014] 13
Figure 11: Many syntactic variations need robust matching approach. [Neves, 2014] 14

Figure 12: Vector Space Model. [6] 16
Figure 13: Relatedness of Vectors. [6] 17
Figure 14: Continuous Bag of Words (CBOW) Model. [6] 19
Figure 15: Skip Gram Model. [6] 19
Figure 16: Translation of English words to Chinese. [Girish] 20
Figure 17: Table contains relations between every two words. [Girish] 21
Figure 18: Classifying of words over their cousins distance. [Girish] 21
Figure 19: Neural Network Architecture. [8] 22
Figure 20: Sentences to vectors. [8] 23
Figure 21: A CBOW structure with Hierarchical Softmax. [12] 25
Figure 22: Negative sampling. [9] 26
Figure 23: The function of cosine. [8] 27
Figure 24: Comparing word vectors. [8] 27
Figure 25: Architecture ofWord2Vec. [Xin Rong, 2018] 28

Figure 26: Steps of Natural Language Processing (NLP). [13] 31
Figure 27: Letter forms. [Mona Diab & Nizar Habash, 2014] 34
Figure 28: letter marks. [Mona Diab & Nizar Habash, 2014] 35
Figure 29: letters (form + mark) in Distinctive format. [Mona Diab & Nizar Habash, 2014] 35
Figure 30: letters (form + mark) in non-Distinctive format. [Mona Diab & Nizar Habash, 2014] 35
Figure 31: Combination example. [Mona Diab & Nizar Habash, 2014] 36
Figure 32: Numbers written format. [Mona Diab & Nizar Habash, 2014] 36

Figure 33: General architecture of our proposed approach. 51
Figure 34: Questions reformulation process steps. 54
Figure 35: Similar words of "العوائد" extracting by CBOW and Skip-Gram models of Word2Vec. 56
Figure 36: Extraction Architecture the results from IRS. 59
Figure 37: Manual test to extract relevant and irrelevant links. 60
Figure 38: Architecture of the Comparison between the results to see if there is improvement. 60

Figure 39: Types of questions use. 67
Figure 40: Example for the result of Farasa Dataset. 69
Figure 41: Result of initial queries (without reformulation). 77
Figure 42: Result of relevant answers by CBOW for test 02. 79
Figure 43: Result of relevant answers by Skip-Gram for test 02. 80
Figure 44: Result of relevant answers by CBOW for test 03. 83
Figure 45: Result of relevant answers by Skip-Gram for test 03. 84
Figure 46: Result of relevant answers by CBOW for test 04. 87
Figure 47: Result of relevant answers by Skip-Gram for test 04. 87
Figure 48: Result of comparison between test04 and test 05 by CBOW. 90
Figure 49: Result of comparison between test 04 and test 05 by Skip-Gram. 90
Figure 50: Query writing window. 99
Figure 51: Window of choice the steps of Questions reformulation process used. 100
Figure 52: window of result. 100
Figure 53: Similar Words Function. 118

List of tables

Table 1: A set of optional diacritics. [Mona Diab & Nizar Habash, 2014].....35
Table 2: Type of systems enumeration symbols. [Mona Diab & Nizar Habash, 2014].....36
Table 3: Works on Deep learning for Arabic NLP Summarization. 47

Table 4: Presentation of the question example type. 53

Table 5: Questions of Benajiba Dataset.....68
Table 6: WH Question type of Benajiba Dataset.....68
Table 7: Example of Question part from the used Dataset Questions. 69
Table 8: Table of Tests. 76
Table 9: Percentage of relevant answers for CBOW model and their justification (Test 02). 81
Table 10: Percentage of relevant answers for Skip-Gram model and their justification (Test 02). 82
Table 11: Percentage of relevant answers for CBOW model and their justification (Test 03). 85
Table 12: Percentage of relevant answers for Skip-Gram model and their justification (Test 03). 86
Table 13: Percentage of relevant answers for CBOW model and their justification (Test 04). 88
Table 14: Percentage of relevant answers for Skip-Gram model and their justification (Test 04). 89
Table 15: Percentage of relevant answers for CBOW model and their justification (Test 05). 91
Table 16: Percentage of relevant answers for Skip-Gram model and their justification (Test 05). 92
Table 17: Result of the comparison. 92
Table 18: Comparison of Named Entities between Test4 and Test5. 93
Table 19: Comparison between percentage of relevant answers for test 4 and 5. 94
Table 20: Number of questions reformulated by each type. 97
Table 21: Table of final results. 98
Table 22: Table of Dataset Questions. 118

LIST OF ACRONYMS

ANN: Artificial Neural Networks.

ANLP: Arabic Naturel Language Processing.

ASR: Automatic Speech Recognition.

AWN: Arabic WordNet.

ATM: Automatic Machine Translation.

API: Application programming Interfaces.

CNN: Convolutional Neural Network.

CA: Classical Arabic.

CBOW: Continuous Bag of Words.

CVSI: Video Script Identification Competition.

DL: Deep Learning.

DBN: Deep Belief Network.

DL4J: Deep Learning4j.

DA: Dialect Arabic.

DR: Detection Rate.

DNN: Deep Neural Network.

FCM: Fuzzy C-Means (Clustering).

IBM: International Business Machines.

Int: Integer (Number).

IRS: Information Retrieval System.

IR: Information Retrieval.

LM: Language Modeling.

ML: Machine Learning.

MRR: Mean Reciprocal Rank.

MSA: Modern Standard Arabic.

NER: Named Entity Recognition.

NLP: Naturel Language Processing.

NLU: Naturel Language Understanding.

NFD: Normalization Form for each character.

NFC: Normalization Form for each character.

NFKD: Normalization Form for each character.

NFKC: Normalization Form for each character.

OCR: Optical Character Recognition.

PANDA: Platform for Architecture-Neutral Dynamic Analysis.

QA: Question Answering.

QE: Query Expansion.

RNN: Recurrent Neural Network.

Re: Regular Expressions.

SVM: Support Vector Machine.

Str: String.

SA: Sentiment Analysis.

TC: Text Categorization.

UCD: Unicode Character Database.

UC: Unix Consultant.

W2V: Word2Vec.

General Introduction

Information science is the discipline which investigates the properties and behavior of information, the forces governing the flow of information, and the means of processing information for optimum accessibility and usability. It is concerned with the body of knowledge relating to the origination, collection, organization, storage, retrieval, interpretation, transmission, and utilization of information.

The search for information can be done through written sources such as photocopied drawings or documents as well as previously extracted results from information or using the Internet

Information technology today, has invaded every area of our daily lives, this has attracted the attention of many researchers in the field of computer science and natural language automatic processing (NLP), to improve are research.

According to research done in Domain of Natural Language Processing (NLP), there are some famous researches, from this researches that are: question answering system and word2vec algorithm that are also the principles of our work.

In the Arab world, these techniques have been used but in English or French, the problem is that it has been limited to some works in Arabic (not rich), although the progress has been in recent years and still incomplete.

In this repport, we propose an approach for improving questions in Arabic language, from a dataset created by benajiba to make the test.

This approach consists of two phases the first is to find the answers for each question exist. And for the second, we did the preprocessing of all the questions with the use of the word2vec algorithm after we find the answers too, and a part of experimentation to compare if there is an improvement between the relevant answers obtained after pretreatment and relevant responses in the original case, and that to measure the performance of our proposed approach.

Problems and Objectives

This project presents the use of the word2vec algorithm with its two forms (CBOW and Skip-Gram) on the questions answering system, the objective is to develope an approach capable of improving the questions, in Arabic language, to find more precise answers.

The word2vec techniques, consists of extracting the vectors as it presents the texts as vector format, and similar words since large texts, most of the research done on this subject in English language, because it is easy to process and analyze . In our case, the language treated is Arabic, this latest which is characterized by a complex morphology, presents one of the great challenges that we have to face.

After choosing the dataset used we don’t know before doing the work if this Dataset will give us a good improvement (Work without knowing the result).

In this repport, we present an approach to improve questions exists in the dataset, to realize this improvement we have based on techniques like, pretreatment and word2vec for the reformulation of the questions.

And in the end we have a test between percentages of relevant answers for original question, and the percentage of relevant answers for questions without reformulation.

Dissertation organization

In our work, we are interested in a modern special method, word2vec, and we also apply word2vec on questions-and-answers with the Arabic language. Word2vec can apply on a large text, or small.

Depending on the nature of our project, we organized our repport into 5 chapters as follows:

The **first chapter** is: the state of the art, in which we explain two generalities, we are talking in a short path of deep learning, question answering system.

In the **second chapter**, we explain in detail word2vec, point after point, Definition, how it works, etc.

The **third chapter** we discuss in detail about natural language processing, after Arabic natural Arabic language, and in the end a part on deep learning for the Arabic NLP.

The **fourth chapter** is devoted to explaining our approach to the application of the word2vec algorithm to improve queries on the Arabic Question answering system.

Explaining all the techniques used to succeed in our approach and apply word2vec on our collection of questions.

In this chapter we have discussed, among others, the following points:

What are the steps to have the reformulated questions?

The approach used to improve the question answering in Arabic language.

In the **fifth chapter**, we present the implementation of our system, explaining also the development tools, and the operating steps of the system illustrated by several tests explained step by step, as well as its results.

Chapter I: State Of The Art

1. Introduction:

With the development of high-speed information technology in various fields, especially informatics, and the human need to adapt to this speed, it was necessary to develop and innovate new ways, and this evolution appears in many of the ideas and research in the field of informatics, for today in this chapter some of them will be introduced. The purpose of this chapter is to give a comprehensive and detailed idea on many concepts, so it is based on two concepts: Deep Learning, Question-Answering System.

For Deep Learning, some information about the concept of deep learning in the form of simplified definitions, types, and how it works are provided, in addition to application areas. Second, the Questions Answering System, their types as well as their applications and then responding to the structure of the system are identified.

2. Deep Learning:

2.1. Deep Learning Definition:

Deep learning is the sub-field of Machine Learning which is the sub-field of Artificial Intelligence [1] deep learning is a set of machine learning algorithms that model high-level abstractions in data using architectures consisting of multiple nonlinear transformations.

A deep learning technology is based on artificial neural networks (ANNs). These ANNs constantly receive learning algorithms and continuously growing amounts of data to increase the efficiency of training processes. The larger data volumes are, the more efficient this process is. The training process is called «deep», because, with the time passing, a neural network covers a growing number of levels. The «deeper» this network penetrates, the higher its productivity is. [2]

2.2. Difference between Neural Network and Deep Learning Neural Network:

Neural Network is a network that can use any network such as feed forward or recurrent network having 1 or 2 hidden layers. But, when the number of hidden layers increases i.e. more than 2 than that, it is known as Deep Learning Neural Network. Neural Network is less complex and requires more information about features for performing feature selection and feature engineering method. On the other hand, Deep Learning Neural Network does not require any information about features rather they perform optimum model tuning and model selection on their own. [3]

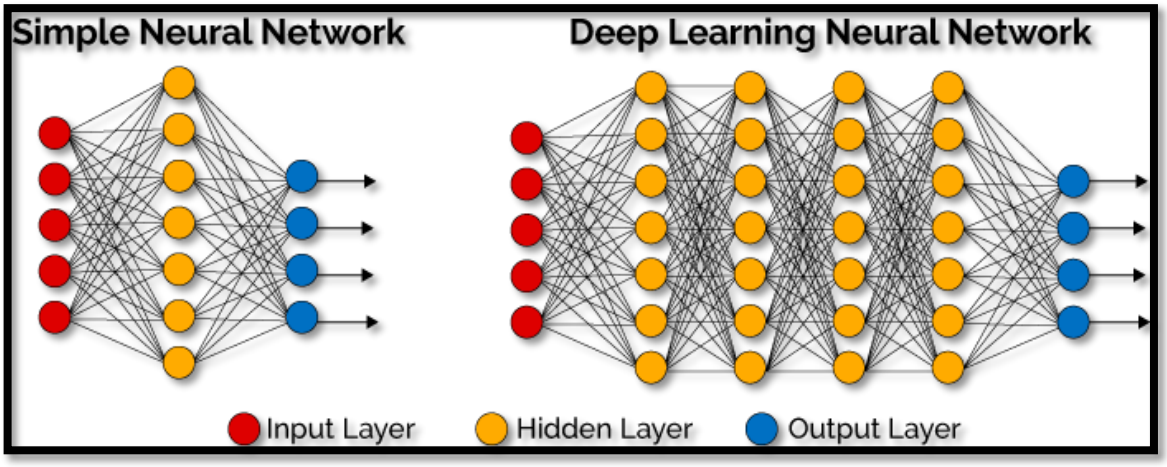


Figure 1: Architecture of Simple Neural Network and Deep Learning Neural Network. [3]

2.3. What Is the Difference between Deep Learning and Machine Learning?:

Deep learning is a kind of traditional machine learning. Classical machine learning is the extraction of new knowledge from a large data array loaded into the machine. Users formulate the machine training rules and correct errors made by a machine. This approach eliminates a negative overtraining effect frequently appearing in deep learning.

In machine learning, users provide a machine with both examples and training data to help the system make correct decisions. This principle is called supervised learning. In other words, in classical machine learning, a computer solves a large number of tasks, but it cannot form such tasks without a human control.

Diversity between machine learning (ML) and deep learning (DL):

- DL requires a lot of unlabeled training data to make concise conclusions while ML can use small data amounts provided by users.
- Unlike ML, DL needs high-performance hardware.
- ML requires features to be accurately identified by users while DL creates new features by itself.
- ML divides tasks into small pieces and then combine received results into one conclusion while DL solves the problem on the end-to-end basis.
- In comparison with ML, DL needs much more time to train.
- Unlike DL, ML can provide enough transparency for its decisions.

The concept of deep learning implies that the machine creates its functionality by itself as long as it is possible at the current time. To infer, deep learning applications use a hierarchical approach involving determining the most important characteristics to compare. [2]

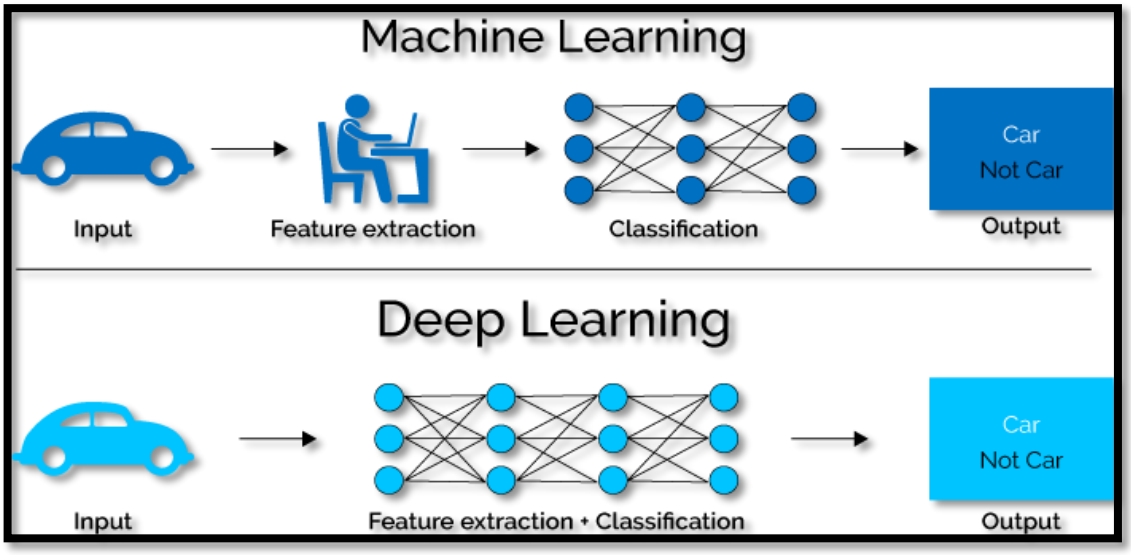


Figure 2: Difference between Machine Learning and Deep Learning. [2]

2.4. Deep Learning type:

In this section, some types of Deep Learning are presented:

2.4.1. Deep Learning for Unsupervised Learning:

This type of deep learning is used when labels of the target variable are not provided and higher correlation has to be computed from observed units for pattern analysis.

2.4.2. Convolutional Neural Networks:

This is a type of network that constitutes of learning weight and biases. Every input layer is composed of a set of neurons where at every input a dot product is performed and move further with the concept of non-linearity. It is a kind of fully-connected type of network that uses SVM/Softmax function as a loss function.

2.4.3. Restricted Boltzmann Machine:

It is a kind of stochastic neural network that consists of one layer of visible units, one layer of hidden units and a bias unit. The architecture is developed in such a way that each visible unit is connected to all hidden units and bias units are connected to all visible and hidden units. During the learning process, the restriction is developed so that no visible unit is connected with any visible units and no hidden unit is connected with any hidden unit.

2.4.4. Recursive Neural Network:

It is the type of deep learning neural network that uses same weights recursively for performing structure prediction about the problem. The stochastic gradient is used for training the network using back propagation algorithm. [3]

2.4.5. Recurrent neural network (RNN):

Is a class of neural networks that includes weighted connections within a layer (compared with traditional feed-forward networks, where connects feed only to subsequent layers). Because RNNs include loops, they can store information while processing new input. This memory makes them ideal for processing tasks where prior inputs must be considered (such as time-series data). For this reason, current deep learning networks are based on RNNs. This tutorial explores the ideas behind RNNs and implements one from scratch for series data prediction. [4]

2.4.6. Autoencoders:

A class of unsupervised learning algorithms in which the output shape is the same as the input that allows the network to better learns basic representations. [Valentino Zocca et al, 2017]

2.4.7. Deep Belief Networks (DBN):

A stochastic generative model where the top layers have symmetric connections between them (undirected, unlike feed-forward networks), while the bottom layers receive the processed information from directed connections from the layers above them. [Valentino Zocca et al, 2017]

2.5. Deep learning application and examples:

Deep learning is currently being utilized in almost every industry starting from Automobile, Aerospace, and Automation to Medical. Here are some of the examples.

- Google, Netflix, and Amazon: Google uses it in its voice and image recognition algorithms. Netflix and Amazon also use deep learning to decide what you want to watch or buy next
- Driving without a driver: Researchers are utilizing deep learning networks to automatically detect objects such as stop signs and traffic lights. Deep learning is also used to detect pedestrians, which helps decrease accidents.
- Aerospace and Defense: Deep learning is used to identify objects from satellites that locate areas of interest, and identify safe or unsafe zones for troops.
- Facebook automatically finds and tags friends in your photos. Skype can translate spoken communications in real-time and pretty accurately too.
- Medical Research: Medical researchers are using deep learning to automatically detect cancer cells
- Industrial Automation: Deep learning is helping to improve worker safety around heavy machinery by automatically detecting when people or objects are within an unsafe distance of machines.
- Electronics: Deep learning is being used in automated hearing and speech translation. [5]

3. Question Answering System:

3.1. Question answering (QA) System Definition:

A question answering (QA) system aims at automatically finding concise answers to arbitrary questions phrased in natural language. It delivers only the requested information, unlike search engines which refer to full documents. [Dong et al, 2011] based on a corpus. [Carlos Amaral & Dominique Laurent, 2006]

3.2. Question answering (QA) system History:

3.2.1. BASEBALL [Green et al., 1963]:

- One of the earliest question answering systems
- Developed to answer users' questions about dates, locations, and the results of baseball matches

3.2.2. LUNAR [Woods, 1977]:

- Developed to answer natural language questions about the geological analysis of rocks returned by the Apollo moon missions
- Able to answer 90% of questions in its domain posed by people not trained on the system

3.2.3. STUDENT:

- Built to answer high-school students' questions about algebraic exercises

3.2.4. PHLIQA:

- Developed to answer the user's questions about European computer systems

3.2.5. UC (Unix Consultant):

- Answered questions about the Unix operating system

3.2.6. LILOG:

- Was able to answer questions about tourism information of cities in Germany [Dr. Mariana Neves,2014]

3.2.7. Open-domain QA: START [Katz, 1997]:

- Utilized a knowledge-base to answer the user's questions
- The knowledge-base was first created automatically from unstructured Internet data
- Then it was used to answer natural language questions [Momtazi, 2011]

3.2.8. WolframAlpha 2009:

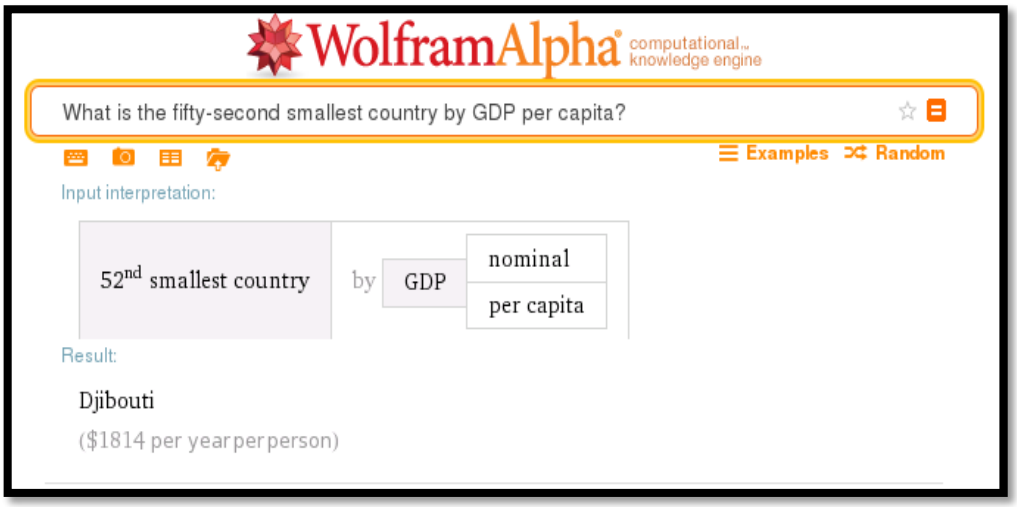


Figure 3: WolframAlpha Question Answering (QA) System. [Lehmann, 2016]

3.2.9. IBM Watson [1] wins at Jeopardy 2011:



Figure 4: Competition between Jeopardy against human experts. [Lehmann, 2016]

- Developed by Deep QA research team at IBM

- Won Jeopardy against human experts
- Massively parallel, runs on a supercomputer
- Industry project: some algorithms are published, but not open source
- Brmson1 platform approximates IBM Watson using open source technology

QA engine Yoda QA2 inspired by Watson runs on a normal server PC [Lehmann, 2016]

3.3. Question answering (QA) system Types:

3.3.1. Open-domain: domain independent QA systems can answer any query from any corpus:

- covers wide range of queries
- low accuracy

3.3.2. Closed-domain: domain specific QA systems are limited to specific domains:

- High accuracy
- limited coverage over the possible queries
- Needs domain expert [Shekarpour, 2015]

3.4. Difference between QAS and search engine:

The most salient difference is that:

- Search engine relies on the human subject to pose queries, not questions.
- A query is a specially formatted string that contains keywords and search engine commands.
- The output of a search engine is typically a set of documents that match the user's query.
  - A Question Answering system on the other hand, takes a natural language question as input.
  - typically returns a specific answer as its output[Wei Kor, 2005]

3.5. Question answering (QA) system Applications:

3.5.1. Personal assistants:

- Personal assistants on smartphones with voice interface
- Need small memory, CPU and battery footprint
- Calendar events
- Weather
- Route planning



Figure 5: Type of Personal assistants (Apple Siri, Google Now, and Window Cortana). [Lehmann, 2016]

3.5.2. Access to Encyclopedic Knowledge:

- "The world's knowledge in your pocket" World knowledge available in DBpedia, Wikidata etc.
- Used by many SQA systems, such as ISOFT, CASIA, Intui, TBSL, AskNow, SINA
- Basis of many QALD evaluation initiative tasks

3.5.3. Smart Home and Entertainment:

- Will get increasingly complex! People will ask devices more complex questions
- Networked devices (Internet of Things)



Figure 6: Smart Home Control. [Lehmann, 2016]

3.5.4. Search:

- QA is becoming part of main stream search engines
- Google:
- Voice Search in 2008
- Knowledge Graph in 2012
- Question Intent Understanding in 2015
- Google can understand superlatives, ordered items, time e.g. "Who was the U.S. President when the Angels won the World Series?" [Lehmann, 2016]

3.6. Question answering (QA) system Architecture:

In order to introduce the domain of answering to the questions, we illustrate the general steps of work that was presented by: "Dr. Mariana Neves" of the Potsdam University entitled "Answering Question", carried out on June 25, 2014, this work is based on the work of "Dr. Saeedeh Momtazi", in the first place, it begins by giving a diagram (figure) contains a real example, after it gives the details of each phase.

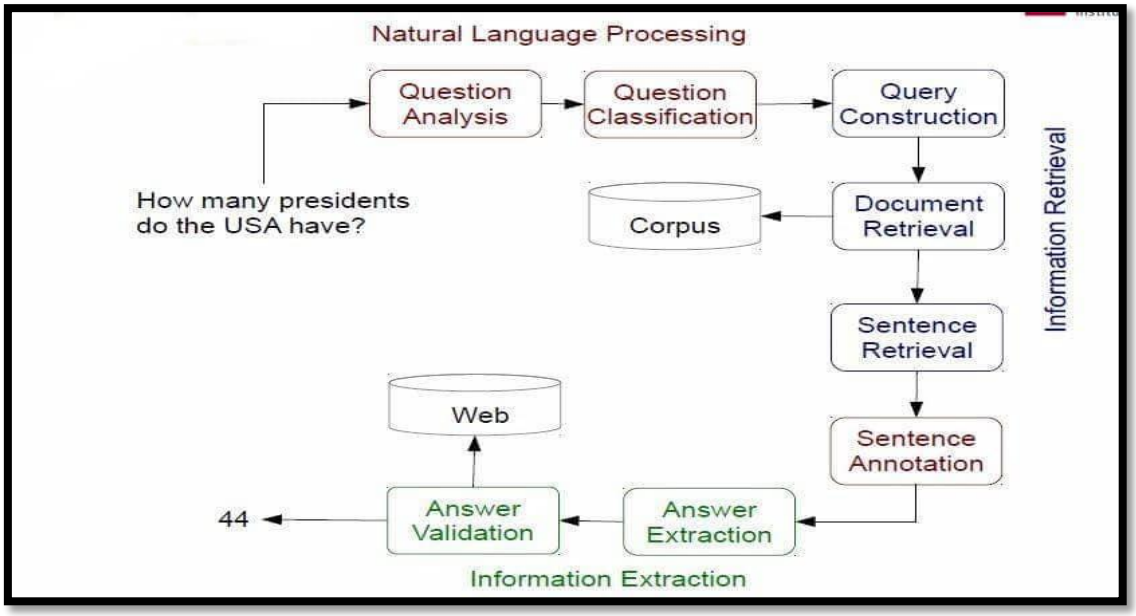


Figure 7: Question answering (QA) system Architecture. [Neves, 2014]

First, post a question has a relation to corpus to begin with:

3.6.1. Questions Categories:

A question is an expression to request information. [Shiqing Hu, 2013]

- **Factoid queries**: WH questions like when, who, where.
- **Yes/ No queries**: Is Berlin capital of Germany?
- **Definition queries**: what is leukemia?
- **Cause/consequence queries**: How, Why, What. What are the consequences of the Iraq war?
- **Procedural queries**: which are the steps for getting a Master degree?
- **Comparative queries**: what are the differences between the model A and B?
- **Queries with examples**: list of hard disks similar to hard disk X.
- **Queries about opinion**: What is the opinion of the majority of Americans about the Iraq war? [Shekarpour, 2015]
- **Indirect Requests** (*I would like you to list …*)
- **Commands** (*Name all the presidents …*)
- **Summaries** (*What are the arguments for and against…*) [Magnini, 2005]

3.6.2. Natural Language Processing:

3.6.2.1. Question Analysis:

- Named Entity Recognition
- Surface Text Pattern Learning
- Syntactic Parsing
- Semantic Role Labeling

3.6.2.1.1. Question Analysis: Named Entity Recognition:

- recognizing the named entities in the text to extract the target of the question

- using the question's target in the query construction step
- Example:
  - Question: "In what country was Albert Einstein born?"
  - Target: "Albert Einstein"

3.6.2.1.2. Question Analysis: Pattern Learning:

- extracting a pattern from the question
- matching the pattern with a list of pre-defined question patterns
- finding the corresponding answer pattern
- realizing the position of the answer in the sentence in the answer extraction step
- Example:
  - Question: "In what country was Albert Einstein born?"
  - Question Pattern: "In what country was X born?"
  - Answer Pattern: "X was born in Y."

3.6.2.1.3. Question Analysis: Syntactic Parsing:

- using a dependency parser to extract the syntactic relations between question terms
- using the dependency relation paths between question terms to extract the correct answer in the answer extraction step

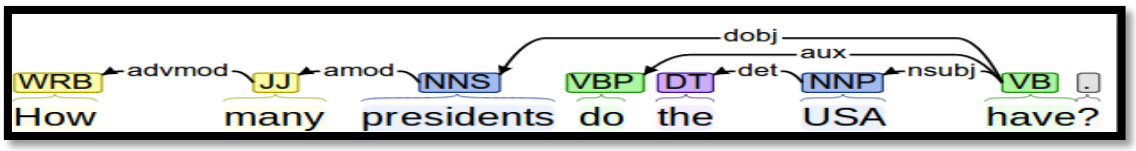


Figure 8: Syntactic Parsing model. [Neves, 2014]

3.6.2.1.4. Question Analysis: Semantic Role Labeling:

- FrameNet: a lexical database for English
- More than 170,000 manually annotated sentences
- Frame Semantics: describes the type of event, relation, or entity and the participants in it.
- Example:
  - John[Cook] grills a fish[Food] on an open fire [Heating-Instrument]
- Frame assignment
- Role labeling
- Jim [Driver] flew his plane [Vehicle] to Texas [Goal] OPERATE−VEHICLE

- Alice [Destroyer] destroys the item [Undergoer] with a plane [Instrument] DESTROYING

- finding the question's head verb
- Example:
  - "Who [Buyer] purchased YouTube [Goods]? "
- COMMERCE−BUY
  - Buyer [Subj, NP] verb Goods [Obj, NP]
  - Buyer [Subj, NP] verb Goods [Obj, NP] Seller [Dep, PP-from]
  - Goods [Subj, NP] verb Buyer [Dep, PP-by]
  - ...
- Example:
  - "In 2006, YouTube [Goods] was purchased by Google [Buyer] for $1.65 billion. "

3.6.2.2. Question Classification:

- classifying the input question into a set of question types
- mapping question types to the available named entity labels
- Finding strings that have the same type as the input question in the answer extraction step
- Example:

- Question: “In what country was Albert Einstein born?”
- Type: LOCATION - Country

● Example (NER):
- S1: “Albert Einstein was born in 14 March 1879.”
- S2: “Albert Einstein was born in Germany.”
- S3: “Albert Einstein was born in a Jewish family.”

● Classification taxonomies
- BBN
- Pasca & Harabagiu
- Li & Roth

● 6 coarse- and 50 fine-grained classes
- ABBREVIATION
- ENTITY
- DESCRIPTION
- HUMAN
- LOCATION
- NUMERIC

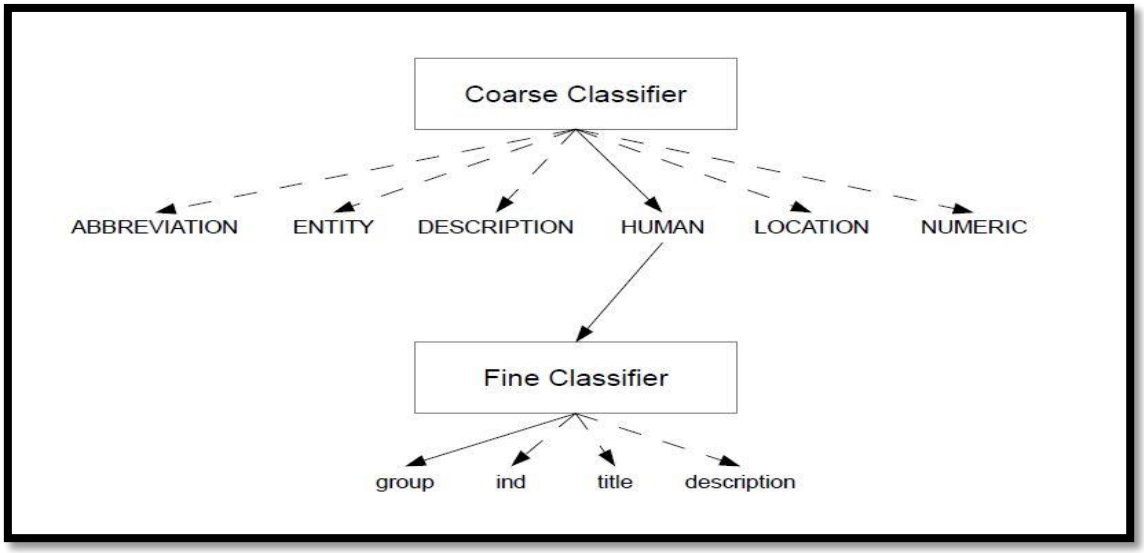


Figure 9: Type of Classification. [Neves, 2014]

● using any kinds of supervised classifiers
- K Nearest Neighbor
- Support Vector Machines
- Naïve Bayes
- Maximum Entropy
- Logistic Regression
- …

● benefiting from available toolkits: Weka, SVM-light, etc.
● considering the confidence measure of the classification to filter the result.

3.6.3. Information Retrieval:

3.6.3.1. Query Construction:

● Goal:
- Formulating a query with a high chance of retrieving relevant documents

● Task:

- Assigning a higher weight to the question's target
- Using query expansion techniques to expand the query

3.6.3.2. Document Retrieval:

- Importance:
  - QA components use computationally intensive algorithms
  - Time complexity of the system strongly depends on the size of the to be processed corpus
- Task:
  - Reducing the search space for the subsequent components
  - Retrieving relevant documents from a large corpus
  - Selecting top n retrieved document for the next steps
- Using available information retrieval models
  - Vector Space Model
  - Probabilistic Model
  - Language Model
- Using available information retrieval toolkits: Lucene, Lemur, etc.

3.6.3.3. Sentence Retrieval:

- Task:
  - Finding small segments of text that contain the answer
- Benefits beyond document retrieval:
  - Documents are very large
  - Documents span different subject areas
  - The relevant information is expressed locally
  - Retrieving sentences simplifies the answer extraction step
- Main problem: sentence brevity
- Information retrieval models for sentence retrieval
  - Vector Space Model
  - Probabilistic Model
  - Language Model
- Jelinek-Mercer Linear Interpolation
- Bayesian Smoothing with Dirichlet Prior
- Absolute Discounting
- Approaches to overcome the sentence brevity problem:
  - Term relationship models
  - Translation model
  - Term clustering model

3.6.3.4. Sentence Annotation (Natural Language Processing):

- Similar to Question Analysis
- annotating relevant sentences using linguistic analysis
  - Named entity recognition
  - Syntactic parsing
  - Semantic role labeling
- Example (NER):
  - S1: "Albert Einstein was born in 14 March 1879."
  - S2: "Albert Einstein was born in Germany."
  - S3: "Albert Einstein was born in a Jewish family."

3.6.4. Information Extraction:

3.6.4.1. Type of Answer:

- **Long answers**, with justification
- **Short answers** (e.g. phrases)
- **Exact answers** (named entities)
- Answer construction:
  - **Extraction**: cut and paste of snippets from the original document(s)
  - **Generation**: from multiple sentences or documents
  - **QA and summarization** (e.g. *what is this story about?)* [Magnini, 2005]

3.6.4.2. Answer Extraction:

- extracting candidate answers based on a various information
  - Question
    - Question Analysis: patterns
    - Question Analysis: syntactic parse
    - Question Analysis: semantic roles
    - Question Classification: question type
- Sentence
  - Sentence Annotation: all annotated data
- using extracted patterns
- Example:
  - Question: "In what country was Albert Einstein born?"
- Question Pattern: In what country was X born?
- Answer Pattern: X was born in Y.
- Example (Pattern):
  - S1: "Albert Einstein was born in 14 March 1879."
  - S2: "Albert Einstein was born in Germany."
  - S3: "Albert Einstein was born in a Jewish family."
- Using question type and entity type
- Example:
  - Question: "In what country was Albert Einstein born?"
  - Question Type: LOCATION - Country
- Example (NER):
  - S1: "Albert Einstein was born in 14 March 1879."
  - S2: "Albert Einstein was born in Germany."
  - S3: "Albert Einstein was born in a Jewish family."
- using syntactic parsing
  - Different wordings possible, but similar syntactic structure

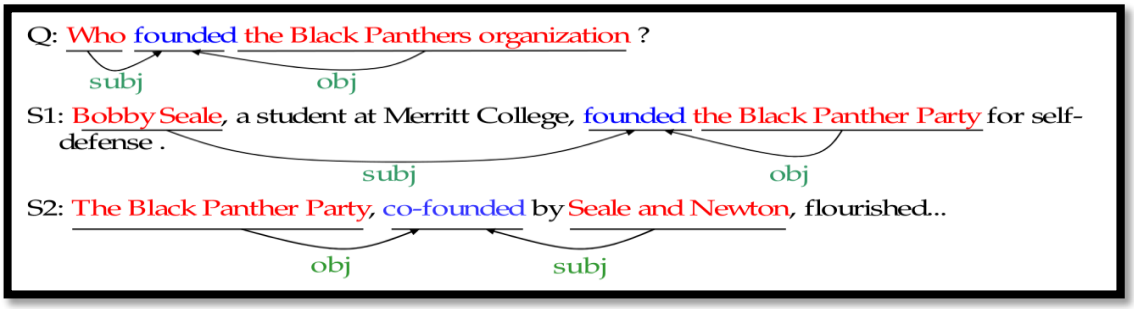


Figure 10: Different wordings possible, similar syntactic structure. [Neves, 2014]

- using syntactic parsing
  - Many syntactic variations need → robust matching approach

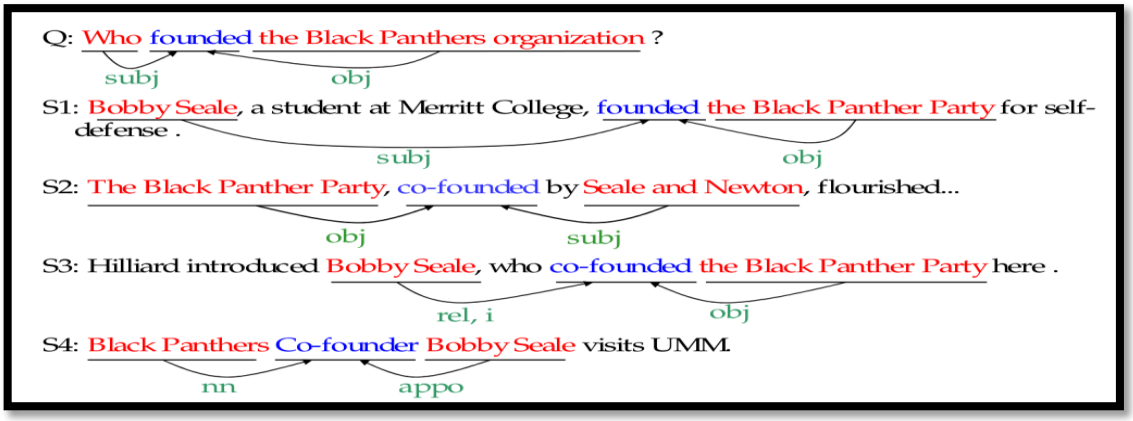


Figure 11: Many syntactic variations need robust matching approach. [Neves, 2014]

- Using semantic roles
- Example:
  - "Who [Buyer] purchased YouTube? [Goods]"
- Example:
  - "In 2006, YouTube [Goods] was purchased by Google [Buyer] for $1.65 billion."

3.6.4.3. Answer Validation:

- Using Web as a knowledge resource for validating answers
- required steps
  - Query creation
  - Answer rating
- Query creation
  - Combining the answer with a subset of the question keywords
  - Choosing different combinations of subsets
- Bag-of-Word
- Noun phrase chunks
- Declarative form
- Answer rating
  - Passing the query to a search engine.
  - Analyzing the result of the search engine.
    - counting the results
    - parsing the result snippets
  - Other possibilities:
    - Using knowledge bases to find relations between the question keywords and the answer [Neves, 2014]

4. Conclusion:

In this first chapter, the fundamental notion of our two concepts was presented: Deep Learning, and Question Answering System, in order to simplify the understanding of the basic concepts on them.

For Deep Learning, different types of Deep Learning are talked about, how he works, and his applications. Therefore, Question Answering System (QA) definition, history, also his types, applications are talked about, then Question Answering System Architecture.

Chapter II: Word2Vec

1. Introduction:

Word embedding techniques are useful for converting text data into vectors with real value, many researchers have tried word2vec, which is demonstrated by the results of the experiments. It has been developed and used for its success. Therefore, it has been used in many fields.

The purpose of this chapter is simply to demonstrate every detail of Word2Vec. This chapter is based on the generality of word2vec and it structure.

2. Word embeddings and word2vec:

2.1. Vector Space Model Representation:

Vector Space Models are used to represent words in a continuous vector space. So the words that are semantically similar are mapped close to each other. Vector Space Models have been used since 1990's for distributional semantics. Some of the older models of estimating continuous representations for words were Latent Semantic Analysis and Latent Dirichlet Allocation. But with time, there have been new and better models for representing and finding similarities between those representations. Whatever the model may be but mostly all of them state that words that appear in the same context share semantic meaning. [6]

2.1.1. Representation:

In Vector Space Model, basically it can represent anything in form of a vector. Say, for example, choose to represent a Sentence as a vector. Then the vector can be a huge N-dimensional vector which represents which words appear in that sentence.

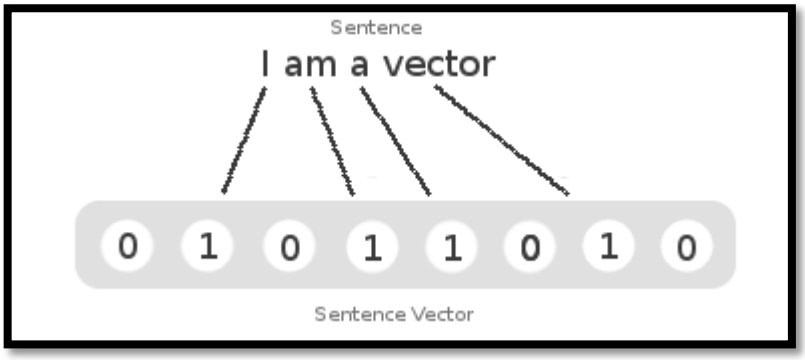


Figure 12: Vector Space Model. [6]

The size of the vector will be the size of vocabulary. In Figure 12, the sentence vector is represented using the words that appear in that sentence. Every word in the vocabulary is given a fixed position in the vector. So suppose if a word "hello" appears in the sentence, then the bit at the position of the word "hello" will be "1".

And this is just one way of representing a sentence. Another way can be to keep a count of words also. For Example, if word "hello" appears twice in a given sentence then this will increment the value at the position of the word "hello" in the vector. Also, can use a vector to know if this word appears in a document of not. So all the elements in the vector will now represent a document, and if any bit is "1" then it means that this word appears in that specific document.

And can make new representations and not just for words, can also represent a word using a vector. Then can use a vector to know if this word appears in a document of not. So all the elements in the vector will now represent a document, and if any bit is "1" then it means that this word appears in that specific document. And can make new representations and not just for words, also can create representations for basically anything like a phrase, document, paragraphs, etc. [6]

2.1.2. Notion of Relatedness:

Relatedness is a property of vector space model to evaluate how much similar to vector are. But in order to find similarity between vectors, the vague notion of relevance has to be converted into a more precise definition that can be implemented with the program analogy. So in this process, a number of assumptions have to be made. These assumptions are based on the representation that has been used.

For example, If the vectors used to represent documents based on the words that appear in that document then one way to find similarity between two documents(or vectors) will be to see how many words those two documents(or vectors) have in common.

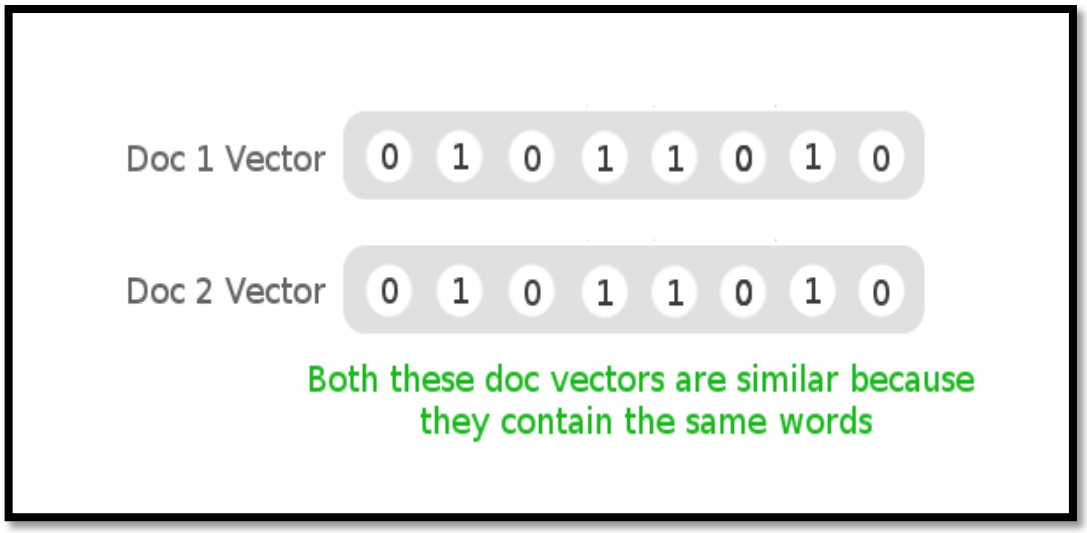


Figure 13: Relatedness of Vectors. [6]

And again this is just one way of finding similarity; you can use different ways to find similarities between two vectors. All these totally depend on your representations. Like for example, you can also use cosine similarity to estimate how much similar two words are in vector space. [6]

2.2. Word Embedding Definition:

Word embeddings are a geometric way to capture the "meaning" of a word via a low-dimensional vector. They are useful in many tasks in Information Retrieval (IR) and Natural Language Processing (NLP), such as answering search queries or translating from one language to another. [7]

2.2.1. Word Embeddings History:

The first time this word 'Word Embeddings' was used in Bengio et al in 2003, where he proposed this new model for distributed representations of words using neural networks. Then in 2008, a paper by Collobert and Weston basically made this 'Word Embeddings' approach mainstream by proposing an 'A unified architecture for Natural Language Processing'. But in 2013, Mikolov published a paper called Word2Vec which basically popularized the use of word embedding and especially pre-trained word embeddings in many NLP tasks. [6]

2.2.2. Why Word Embedding Models Are Used? :

Computer Vision systems work with rich, high dimensional vectors to store and process the data. But this was not the case in Natural Language Processing tasks. In NLP, all the words are treated as a discrete atomic symbols i.e. word 'Cat' may be stored as '123' and word 'Dog' may be stored as '346'. And these numbers carry no semantic or syntactic information. They are just some numbers representing a word. So while processing 'cat' this model can leverage very little about what it knows about 'dogs'. Representing words as unique, discrete ids furthermore leads to data sparsity, and usually, means that more data may be needed in order to successfully train statistical models. [6]

2.2.3. Word Embedding Models:

Word Embedding has two models: Word2Vec and Glove, as our work based on Word2Vec we will talk just about it.

2.3. Word2Vec Definition:

Word2Vec is the most popular word embedding model. Word2Vec is considered as a starter of 'Deep Learning in NLP'. However, Word2Vec is not deep. But the output of Word2Vec is what Deep Learning models can easily understand. Word2vec is basically a computationally efficient predictive model for learning word embeddings from raw text.

The purpose of Word2Vec is to group words that semantically similar in vector space. It computes similarities mathematically. Given huge amount of data, Word2Vec model can create a very rich distributed representation of words that also preserves the semantic relationships that words have with other words. A train Word2Vec model can basically understand the meaning of a word based on the past appearances.

The most famous example with be \(v(king) – v(man) + v(women) ≈ v(queen)\). This means that words can be clustered, find similar words, and find similar relationships of different words and more.

Another good example will be that a \(v(man) \) is related to a \(v(boy)\) in a similar way how \(v(woman)\) is related to a \(v(girl)\). [6]

Or in simply way:

Word2Vec is a machine learning algorithm created by Google in 2013, which makes highly accurate guesses about a word's meaning from raw text. It helps computers understand language with no human supervision. [8]

2.3.1. Continuous Bag of Words (CBOW) Model:

In CBOW model, the input to the neural networks is \(w_(i-2), w_(i-1), w_(i+1), w_(i+2)\) and the word have to be predicted\(w_i\). Basically, it is saying that given the

context predict the word in center. The below image will help you understand the model better.

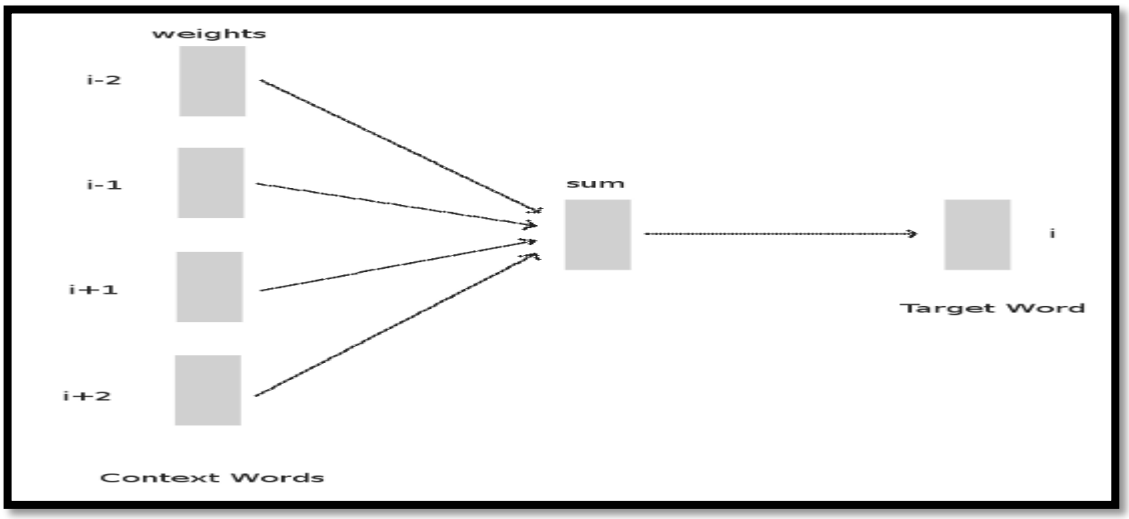


Figure 14: Continuous Bag of Words (CBOW) Model. [6]

The CBOW model gives better accuracy and is faster than Skip gram model but it requires huge of the data. [6]
CBOW predicts target words (e.g. 'mat') from the surrounding context words ('the cat sits on the'). [9]

2.3.2. Skip Gram Model:

The skip gram model is the opposite of CBW model. In this model, the \(w_i\) is the input and the output is \(w_(i-2), w_(i-1), w_(i+1), w_(i+2)\). Basically, it is saying that given a target word, predict the context words. The below image will help you understand better.

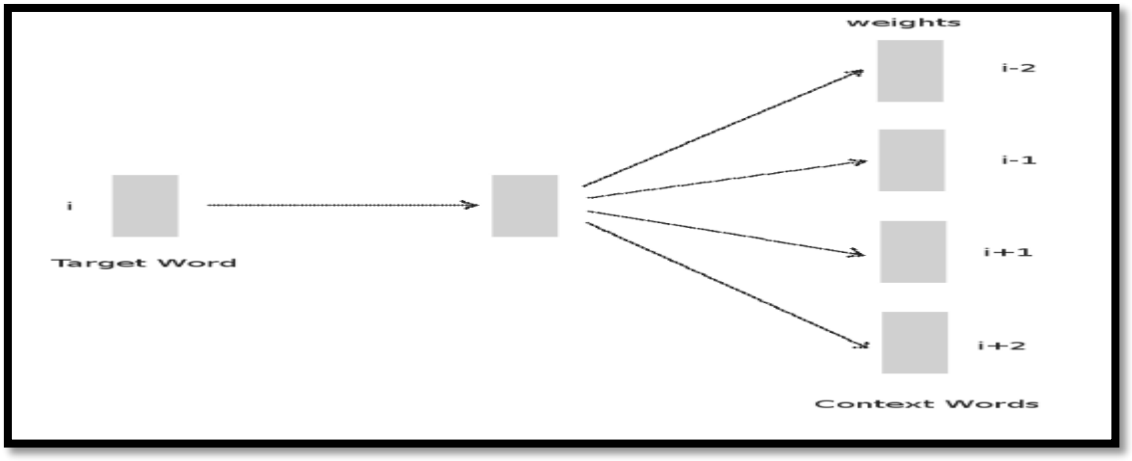


Figure 15: Skip Gram Model. [6]

According to Mikolov, this skip gram model works better in practice because most of the time, people don't have huge amount of data. So skip gram model is better for small training data. [6]

Skip-gram predicts surrounding context words from the target words (inverse of CBOW).

2.4. Word2Vec Applications:

2.4.1. Word Similarity:

Classic Methods: Edit Distance, WordNet, Porter's Stemmer, and Lemmatization using dictionaries.

- Easily identifies similar words and synonyms since they occur in similar contexts.
- Stemming (thought -> think).
- Inflections, Tense forms.
- eg. Think, thought, ponder, pondering,
- eg. Plane, Aircraft, Flight.

2.4.2. Machine Translation:

Classic Methods: Rule-based machine translation, morphological transformation.

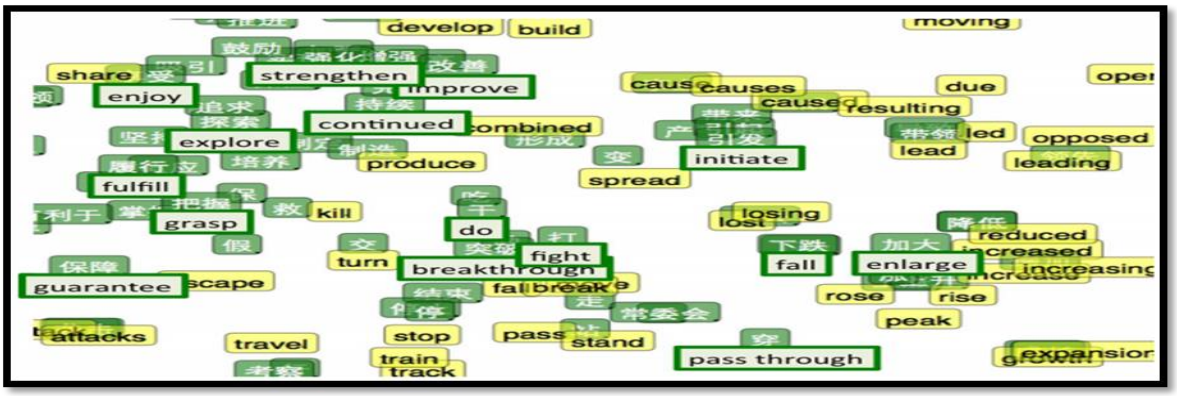


Figure 16: Translation of English words to Chinese. [Girish]

2.4.3. Part-of-Speech and Named Entity Recognition:

Classic Methods: Sequential Models (MEMM, Conditional Random Fields), Logistic Regression.

2.4.4. Relation Extraction:

Classic Methods: OpenIE, Linear programing models, Bootstrapping.

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Figure 17: Table contains relations between every two words. [Girish]

2.4.5. Sentiment Analysis:

Classic Methods: Naive Bayes, Random Forests/SVM.

- Classifying sentences as positive and negative.
- Building sentiment lexicons using seed sentiment sets.
- No need for classifiers, just use cosine distances to compare unseen reviews to known reviews.

```
Enter word or sentence (EXIT to break): sad

Word: sad  Position in vocabulary: 4067

                                                  Word       Cosine distance
--------------------------------------------------------------------------------
                                             saddening                 0.727309
                                                   Sad                 0.661083
                                              saddened                 0.660439
                                         heartbreaking                 0.657351
                                         disheartening                 0.650732
                                          Meny_Friedman                0.648706
                                 parishioner_Pat_Patello               0.647586
                                             saddens_me                0.640712
                                           distressing                 0.639909
                                      reminders_bobbing                0.635772
                                       Turkoman_Shiites                0.635577
                                                saddest                0.634551
                                            unfortunate                0.627209
                                                  sorry                0.619405
                                            bittersweet                0.617521
                                                 tragic                0.611279
                                              regretful                0.603472
```

Figure 18: Classifying of words over their cousins distance. [Girish]

2.4.6. Co-reference Resolution:

- Chaining entity mentions across multiple documents - find and unify the multiple contexts in which mentions occur?

2.4.7. Clustering:

- Words in the same class naturally occur in similar contexts, and this feature vector can directly be used with any conventional clustering algorithms (K-Means, agglomerative, etc.). Human doesn't have to waste time hand-picking useful word features to cluster on.

2.4.8. Semantic Analysis of Documents:

- Build word distributions for various topics, etc. [Girish]

2.5. Word2vec use in our world:

W2V helps solve many difficult real world problems: conversational systems (think chatbots), recommendation algorithms, and knowledge extraction. Many companies use w2v models to interact with their users more effectively.

1. Stitchfix, an online personal styling service for women, has customers fill out a web form requesting specific styles of clothing. They use w2v algorithm to parse through this text and determine when to send a box of vacation clothing for a trip to Belize instead of maternity outfits.

2. Spotify, a music streaming service, uses w2v to find and recommend related artists to users based on the tracks they are streaming. This predicts that a 2pac fan is much more likely to enjoy Notorious B.I.G. than Florence + the Machine.

3. Dice, an online technology job board, uses w2v to create clusters of similar words in job postings. They can infer that job postings containing "data mining" should be clustered with those that contain "statistics" since these words are similar. [8]

2.6. Word2Vec structure:

2.6.1. Define the entry and the exit of Word2Vec (Input/Output):

The input layer for Word2Vec is a large amount of text (corpus) which is processed in the hidden layer.

The output is a list of similar words**.**

As Word2Vec gets more input text, it becomes smarter and gives better results. [8]

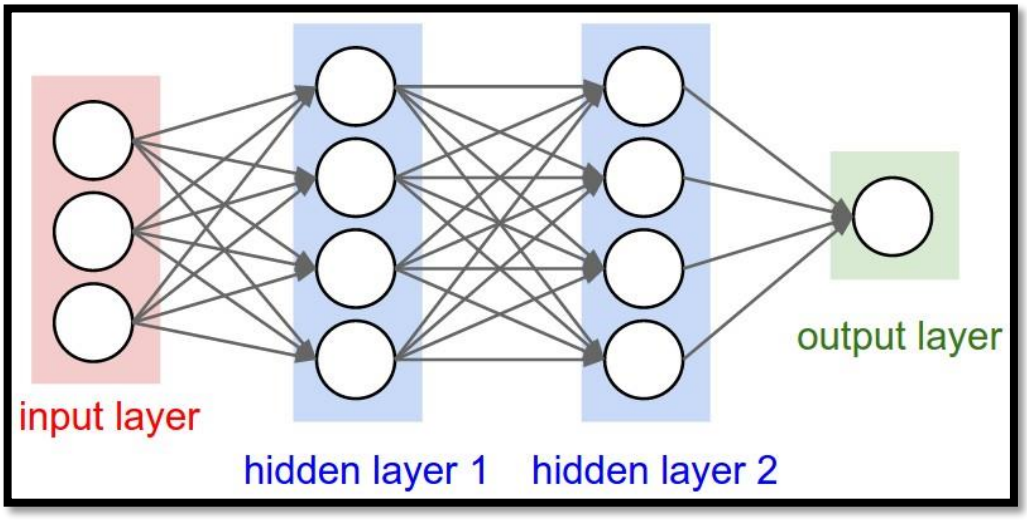


Figure 19: Neural Network Architecture. [8]

2.6.2. Tokenization:

The tokenization of corpus will be done in unigram, bi gram, tri gram.., or divide corpus into words, and set of two words, etc.

2.6.3. Turning Sentences into Vectors and Count Number of existing word in vectors:

In order to be processed by a machine, words must be given a numerical value. W2V takes the text input and turns it into many vectors. In math, a vector is a quantity having a direction and magnitude. Vectors are used to compare a point in space relative to another. Vectors can be added and subtracted to other vectors. They can be manipulated to group words with similar meanings into clusters and produce analogies.

Let's explore the example below:

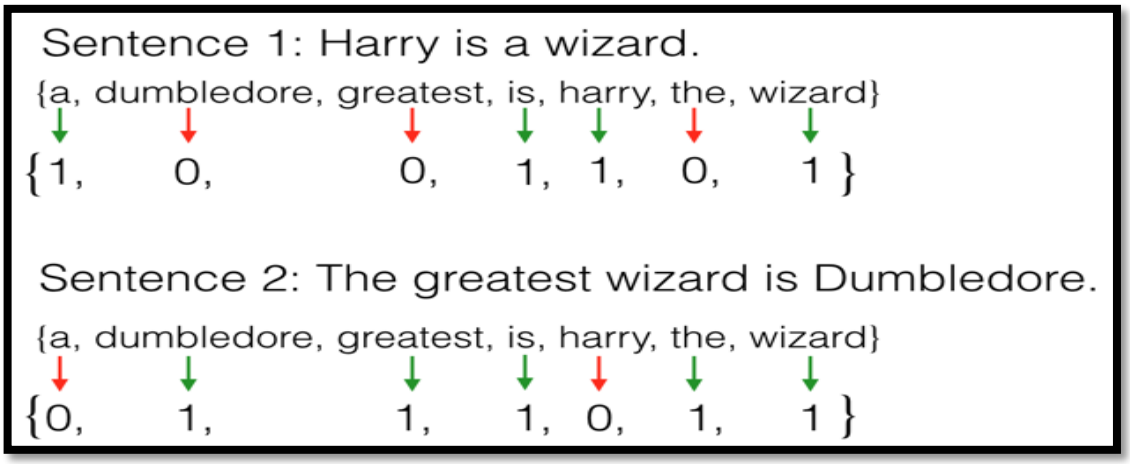


Figure 20: Sentences to vectors. [8]

The lists of unique words, or features, from the 2 sentences are: {a, dumbledore, greatest, is, harry, the, wizard}.

This is used to create feature vectors for each sentence by counting the number of times each word appears in each sentence.

In sentence 1, "dumbledore", "greatest" and "the" appear 0 times, and "a", "is", "harry" and "wizard" appear 1 time each.

In this example, there is a maximum of 7 features since there are only 7 unique words. In real world applications there may be tens of thousands of features (or words) the algorithm must consider. More features generally mean better results, but sometimes reducing the number of features is beneficial for speed improvements. This is called dimensionality reduction.

In text analytics, stop words like "the", "of", and "and" are often removed. However, this is not recommended in w2v since it accepts a list of sentences and removing stop words may alter the meaning of a sentence. [8]

2.6.4. The hidden layer:

2.6.4.1. Choosing from Word2Vec Architecture (CBOW or Skip Gram):

Algorithmically, these models are similar, except that CBOW predicts center words from context words.

While the skip - gram does the inverse and predicts source context - words from the center words.

For example , if the sentence : " The quick brown fox jumps " , then CBOW tries to predict "" brown "" from " the " , " quick " , " fox " , and " jumps " .While skip - gram tries to predict " the " , " quick " , " fox " , and " jumps " from " brown ".

Statistically it has the effect that CBOW smoothes over a lot of the distributional information (by treating an entire context as one observation). For the most part, this turns out to be a useful thing for smaller datasets.

However, skip - gram treats each context - target pair as a new observation and this tends to do better when larger datasets are available. [Chip Huyen & Danijar Hafner, 2012]

2.6.4.2. Training Neural Network:

The neural network is a very small one; let's go over it step by step:

- It has an input layer that receives one-hot encoded words which are of dimension V (the size of the vocabulary).
- It averages them, creating a single input vector.
- That input vector is multiplied by a weights matrix W (that has size VxD, being D nothing less than the dimension of the vectors that you want to create). That gives you as a result a D-dimensional vector.
- The vector is then multiplied by another matrix, this one of size DxV. The result will be a new V-dimensional vector.
- That V-dimensional vector is normalized to make all the entries a number between 0 and 1, and that all of them sum 1, using the softmax function, and that's the output. It has in the i-th position the predicted probability of the i-th word in the vocabulary of being the one in the middle for the given context.

That's it. Oh, are you wondering where the word vectors there are? Well, they are in that W weights-matrix. Remember its size is VxD so the row i is a D-dimensional vector representing the word i in the vocabulary.

In fact, think that if you had just one word used as context, your input vector would be a one-hot encoded representation of that word, and multiplying it by W would be just like selecting the corresponding word vector.

There is a difference that Skip gram is just the same as CBOW, but with one big difference: instead of predicting the word in the middle given all the others, you train to predict all the others given the one in the middle. And it actually works a little better than CBOW.

Apart from the algorithms, word2vec proposed many optimizations to them, such as:

- Give more weight to the closer words in the window.
- Remove rare words (that only appear a few times) from the texts.
- Treating the common word pairs like "New York" as a single word.
- Negative sampling: it's a technique to improve the training time. When you train with one word and its context, you usually update all of the weights in the neural network (remember there a lot of them!). With negative sampling you just update the weights that correspond to the actual word that should have been predicted and some other randomly chosen words (the suggestions is to select between 2 and 20 words), letting most of them as they were. [10]

2.6.5. Training algorithm:

2.6.5.1. Hierarchical Softmax:

Hierarchical softmax has been proposed in the context of language models to speed up training by Morin and Bengio (2005), following prior work by Goodman (2001). The idea is to decompose the softmax layer into a binary tree with the words of the vocabulary at its leaves, such that the probability of a word given a context can be decomposed into probabilities of choosing the correct child at each node along the path from the root node to that leaf. This reduces the number of necessary updates from a linear to a logarithmic term in the vocabulary size. [11]

Word2vec provides two kinds of algorithms in both CBOW and Skip-Gram: Hierarchical Softmax and Negative Sampling.

Hierarchical Softmax is a key technique which greatly improves the performance of Word2vec. It means that the output layer corresponds to a binary tree, or a Huffman tree to be precise. It is built according to the word count. The algorithm traverses down the tree and make binary classification on every non-leaf nodes, and each creates a probability by using the sigmoid function σ, until reaching the leaf nodes corresponding to every word in the vocabulary In addition, Word2vec uses Huffman Code to distinguish the positive examples as 0 and the negative examples as 1. The optimization objective of the model is very similar with that of Logistic Regression, and can be trained using Stochastic Gradient Descent method. [12]

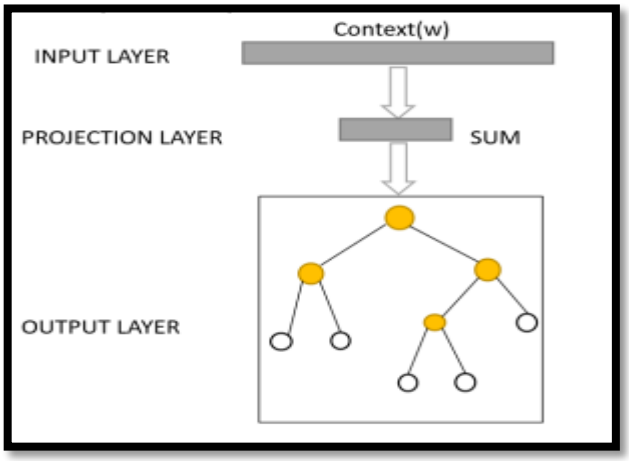


Figure 21: A CBOW structure with Hierarchical Softmax. [12]

2.6.5.2. Negative sampling:

Negative sampling reduces computation by sampling just N negative instances along with the target word instead of sampling the whole vocabulary.

Technically, negative sampling ignores most of the '0' in the one-hot label word vector, and only propagates and updates the weights for the target and a few negative classes which were randomly sampled.
More concretely, negative sampling samples negative instances (words) along with the target word and minimize the log-likelihood of the sampled negative instances while maximizing the log-likelihood of the target word.

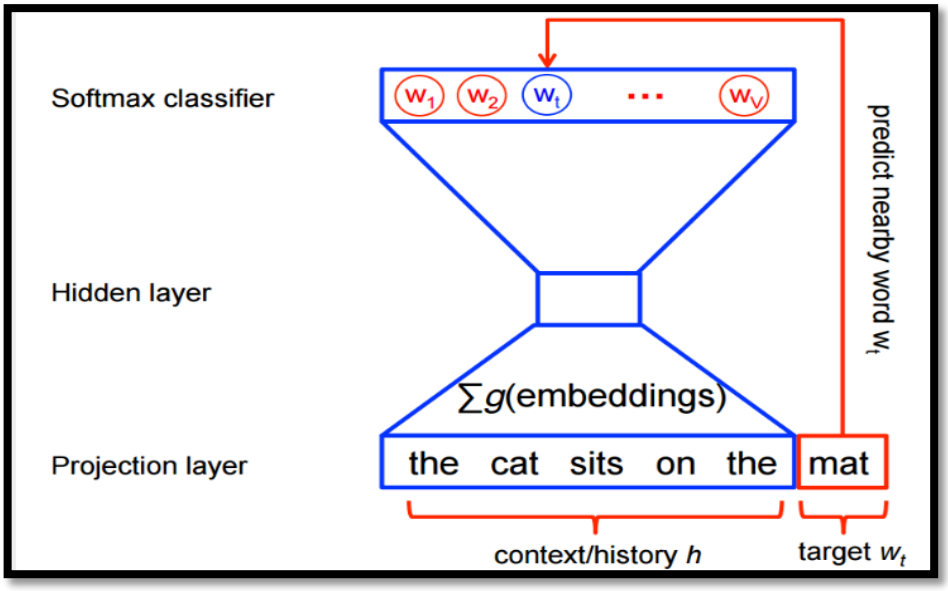


Figure 22: Negative sampling. [9]

2.6.5.2.1. How the negative samples are chosen?:

The negative samples are chosen using a unigram distribution.

Essentially, the probability for selecting a word as a negative sample is related to its frequency, with more frequent words being more likely to be selected as negative samples.

Specifically, each word is given a weight equal to its frequency (word count) raised to the 3/4 power. The probability for a selecting a word is just its weight divided by the sum of weights for all words. [9]

2.6. 6.Final Output of word2vec:

After completing the training, end up with vector representations of the language's vocabulary words.

2.6.7. Drawing the graph:

After turning word into vectors, these vectors can be graphed.

In w2v, vectors are in a high-dimensional space meaning there are many more axes or directions in the space. [8]

2.6.8. Word Similarity:

How does the model know when two words are similar? It uses cosine similarity to compare vectors.

The cosine of an angle (see green below) is the ratio of the adjacent side of a triangle to the hypotenuse.

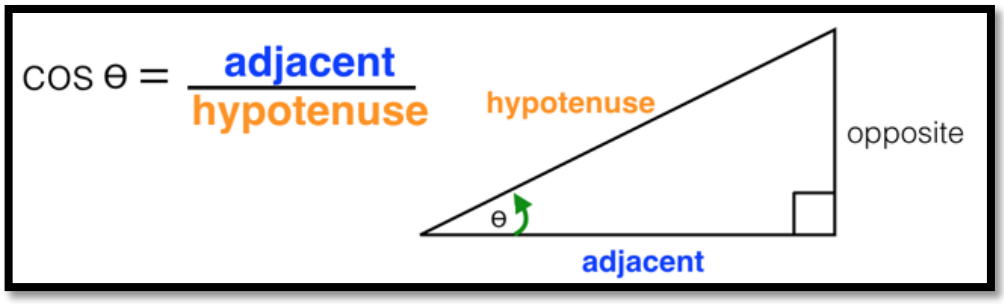


Figure 23: The function of cosine. [8]

This can be used to calculate unknown lengths or angles in triangles. Trigonometry functions are used in electronic communications, signal processing, designing buildings, and many other real-life applications.

After turning word into vectors, these vectors can be graphed. The cosine of the angle between word vectors can be used to compare how similar words are. The smaller an angle is, the larger (closer to 1) the cosine value will be. In this case, a larger cosine value indicates word vectors which have similar meanings.

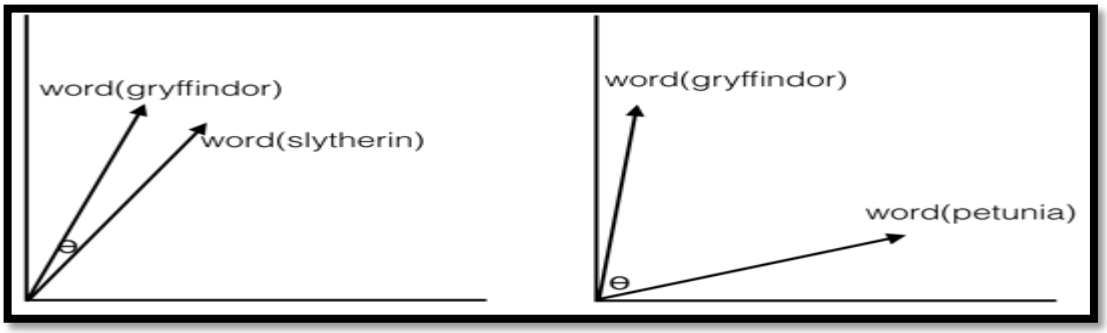


Figure 24: Comparing word vectors. [8]

On the left, the angle between words “gryffindor” and “slytherin” is small, so the cosine of the angle between them is large. This means they are similar. This makes sense- they are both Hogwarts houses. On the right, words “gryffindor” and “petunia” are moving in separate directions. The angle between them is large, so the cosine of the angle between them is small which means these words are not similar.

This is an example in a two-dimensional space with only an X and Y axis. In w2v, vectors are in a high-dimensional space meaning there are many more axes or directions in the space. [8]

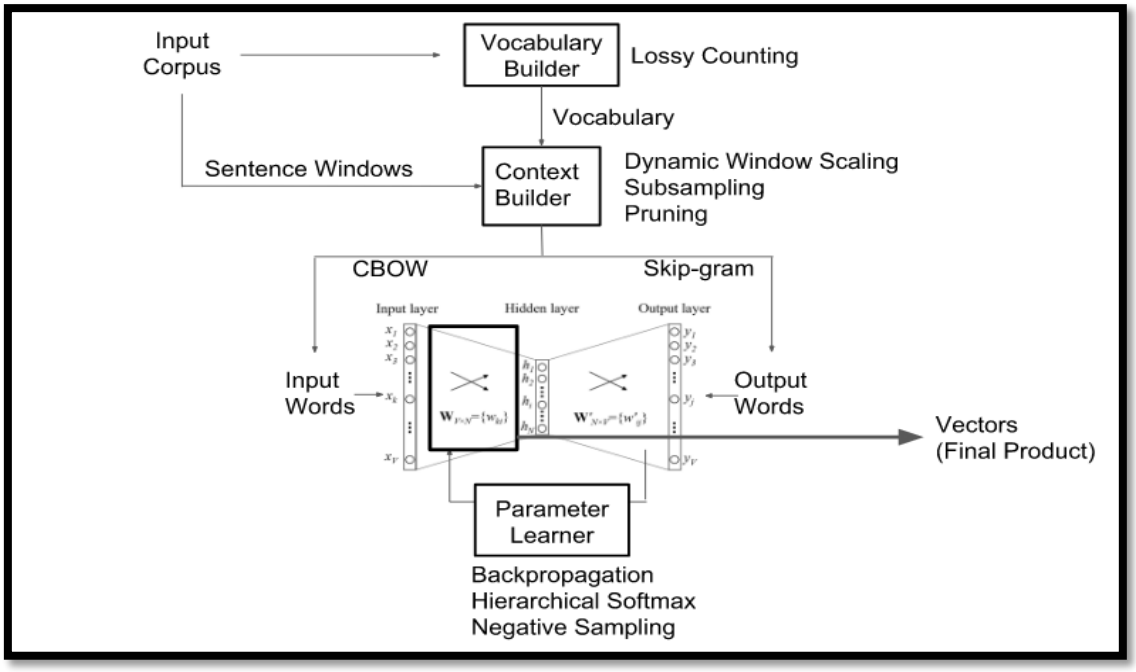


Figure 25: Architecture ofWord2Vec. [Xin Rong, 2018]

3. Conclusion:

In this second chapter, the word embedding and word2vec algorithm were presented, then for Word embedding, the definition, history, models, and the use of word embedding are talked about. Therefore, some information about word2vec: definition, models, and then the word2vec structure are presented in detail.

Chapter III: Deep Learning for Arabic NLP

1. Introduction:

The Arabic language is one of the most widely used languages, such as English, because of its richer morphological structure. It also has many challenges that require special treatment, therefore, Arabic natural language processing has become attractive to researchers due to its complexity and the scarcity of available resources; as a result, the importance of addressing this language has been noted. Nevertheless, deep learning begins to work in some areas of Arabic natural processing.

The purpose of this chapter is to give some idea about Natural Language Processing (NLP), Arabic Natural Language Processing, and Deep Learning for Arabic Natural Language Processing (works), for natural language processing, some information about Natural Language Processing are presented (NLP like a definition, steps, examples, as well as application. Despite, Arabic Natural Language Processing is tackled (NLP a definition of Arabic language, Arabic Script is started by. Therefore, some of work done by using Deep Learning with Arabic Natural Language Processing is given; in each type, one or two method (experiments) and their results are put on.

2. Natural Language Processing (NLP):

2.1. Natural Language Processing (NLP) Definition:

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP researchers aim to gather knowledge on how human beings understand and use language so that suitable tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform desired tasks. [KHUDHUR, 2014]

2.2. Natural Language Processing (NLP) Steps:

There are general five steps:

2.2.1. Lexical Analysis: It involves identifying and analyzing the structure of words. Lexicon of a language means the collection of words and phrases in a language. Lexical analysis is dividing the whole chunk of txt into paragraphs, sentences, and words.

2.2.2. Syntactic Analysis (Parsing): It involves analysis of words in the sentence for grammar and arranging words in a manner that shows the relationship among the words. The sentence such as “The school goes to boy” is rejected by English syntactic analyzer.

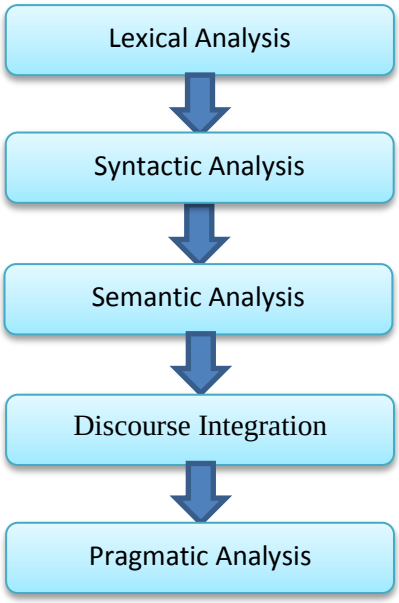


Figure 26: Steps of Natural Language Processing (NLP). [13]

2.2.3. Semantic Analysis: It draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the task domain. The semantic analyzer disregards sentence such as “hot ice-cream”.

2.2.4. Discourse Integration: The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.

2.2.5. Pragmatic Analysis: During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge. [13]

2.3. Natural Language Processing (NLP) Application:

NLP is currently being used in a variety of areas to solve difficult problems.

2.3.1. Machine Learning:

NLP is used in conjunction with machine learning techniques to perform tasks such as emotion detection, sentiment analysis, dialogue act recognition, spam email classification etc. Machine learning techniques require data to train algorithms.

Natural language in the form of tweets, blogs, websites, chats etc. is a huge source of data. NLP plays a very important role in data collection by converting natural language into a format that can be used by machine learning techniques to train algorithms.

Spam classification is a classic example of the use of NLP in machine learning. Spam classification refers to the process of classifying emails as "spam or ham" based on the contents of the email. For instance an email "Hello, congratulations you have won $100000000" can be classified as "spam" and another email "As per our discussion, please find attached minutes of the meeting" can be classified as "ham".

NLP tools and techniques help convert the text of these emails into feature vectors that can be used by machine learning applications to train the algorithm and then predict a new email instance as spam or ham.

2.3.2. Human Computer Interaction:

Human computer interaction has evolved from simple mouse-and-keyboard desktop-based interaction to more natural interaction involving speech and gestures. For example, Amazon's Alexa and Apple's Siri are two of the prime examples of such interaction where humans use speech to interact with the system and perform different tasks. Another example of natural interaction is Google's homepage where you can perform search operations via speech. Natural language processing lays at the foundation of such interaction.

2.3.3. Information Extraction:

Another important NLP task is to extract useful information from documents that can be used for different purposes. For instance, publicly traded companies are required to publish financial information and make it available to their shareholders. NLP can be used to extract financial information from these kinds of documents to automatically gather information on how a company (or industry) is doing.

If mined carefully, the information extracted from these sources can help companies make appropriate decisions. For instance, if a company didn't perform very well last quarter then you could use NLP to determine this and then automatically sell your shares of their stock.

2.3.4. Summarization:

Not every piece of information in a text document is useful, so any "fluff" may be cut out and read only what is important. Summarization refers to condensing a document in a way that it contains only key pieces of information, leaving behind all the waste.

For example, instead of having to read a multi-page news story you could use summarization to extract only the important information about whatever news event the article was about. Or a summary could be made as a preview of the article to help you decide if you want to read the full text or not. [14]

2.3.5. Reduce words to their root (or stem):

Using Porter Stemmer, or break up text into tokens using Tokenizer. **[**15**]**

2.4. Natural Language Processing systems (NLP) Examples in artificial intelligence:

2.4.1. Communication:

Many communication applications such Facebook Messenger are already using artificial intelligence. On the whole, Facebook looks very interested in AI. Some months ago, Facebook announced its M service that promises to become your personal assistant (with the public launch date tbd): “M can do anything a human can.” When you request something that M can’t do on its own, it sends a message to a Facebook worker and, as they work with the software, the AI begins to learn. Another interesting application of natural language processing is Skype Translator, which offers on-the-fly translation to interpret live speech in real time across a number of languages. Skype Translator uses AI to

help facilitate conversation among people who speak different languages. This is great news! Without language barriers, people can communicate using the language they are comfortable with, which will in turn speed up a range of businesses processes.

2.4.2. Faster diagnosis:

Examples of natural language processing systems in artificial intelligence are also in hospitals that use natural language processing to indicate a specific diagnosis from a physician's unstructured notes. For example, NLP software for mammographic imaging and mammogram reports support the extraction and analysis of data for clinical decisions, as a study published in Cancer affirms. The software is able to determine breast cancer risk more efficiently, decrease the need for unnecessary biopsies and facilitate faster treatment through earlier diagnosis. According to the study, artificial intelligence reviewed 500 charts in but a few hours, saving over 500 physician hours. "Accurate review of this many charts would be practically impossible without AI," the author Stephen T. Wong said.

2.4.3. Customer Review:

Natural language processing in artificial intelligence applications makes it easy to gather product reviews from a website and understand what consumers are actually saying as well as their sentiment in reference to a specific product. Companies with a large volume of reviews can actually understand them and use the data collected to recommend new products or services based on customer preferences. This application helps companies discover relevant information for their business, improve customer satisfaction, suggest more relevant products or services and better and understand the customer's needs.

2.4.4. Virtual digital assistants:

Thanks to smartphone, virtual digital assistant (VDA) technologies (automated software applications or platforms that assist the human user by understanding natural language) are currently the most well-known type of artificial intelligence. Many companies understand the importance of the VDAs for their businesses and are investing significant resources to stay up to date. According to a study published by Research and Markets, VDA users will grow from 390 million in 2015 to 1.8 billion worldwide by 2021, while enterprise VDA users will rise from 155 million in 2015 to 843 million over the same period. VDAs are able to assist the consumers with transaction activities or optimize the call center operations to offer a better customer experience and reduce the operational costs. These applications in other devices such as PCs programs will be increasingly seen, smart home systems, automobiles and in the enterprise market. [16]

3. Arabic Natural Language Processing (ANLP):

3.1. Arabic Language Definition:

The Arabic language is considered one of the most widely used languages of the world because there are more than 250 million inhabitants in Arab. Besides, Arabic is the official

language of many countries in the Arab world, including Egypt, Algeria, Iraq, Jordan, Lebanon, Saudi Arabia, Syria, and Kuwait. Arabic is the language of the Quran. [KHUDHUR, 2014]

The Arabic language is comprised of 28 letters (25 consonants and three long vowels).

It is a cursive language, in which words consist of cursive Arabic letters connected to one another. In addition, the writing in Arabic runs from right to left. Like other languages, such as Japanese and Korean, Arabic has no capitalization. Arabic letters have different shapes depending on their position in the word. Unlike English, which has dedicated letters to represent short vowels, Arabic has diacritics that play the same role as short vowels in English and determine the pronunciation or the sound of the letter. All words in Arabic are derived from a root which is composed of constants. These are generally three or four letters called radicals. [Alotaibi, 2015]

3.2. Arabic Language Types:

There are three main types of Arabic; these types are:

3.2.1. Classical Arabic CA: CA is the oldest version of Arabic, which is used in the earliest age of Arab nation.

3.2.2. Modern Standard Arabic MSA: The MSA is the formal Arabic language, which is used nowadays in education, books, newspapers, media, and even as the official language of Arabic countries.

3.2.3. Dialect Arabic DA: DA is a kind of colloquial language that differs from region to region in Arab countries. [Alotaibi, 2015]

3.3. Arabic Language Script:

- An alphabet
- Written right-to-left
- Letters have allography variants
- No concept of "capitalization"
- Optional diacritics
- Common ligatures
- Used to write many languages besides Arabic: Persian, Kurdish, Urdu, Pashto, etc. [Mona Diab & Nizar Habash, 2014]

1. **Alphabet:**
   - Letter forms:



Figure 27: Letter forms. [Mona Diab & Nizar Habash, 2014]

   - letter marks:



Figure 28: letter marks. [Mona Diab & Nizar Habash, 2014]

- letters (form + mark):
  - Distinctive

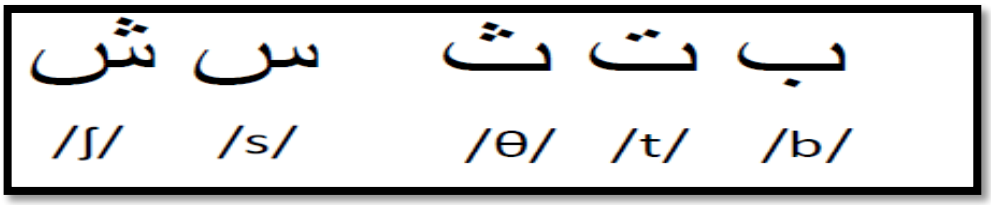


Figure 29: letters (form + mark) in Distinctive format. [Mona Diab & Nizar Habash, 2014]

  - Non-distinctive

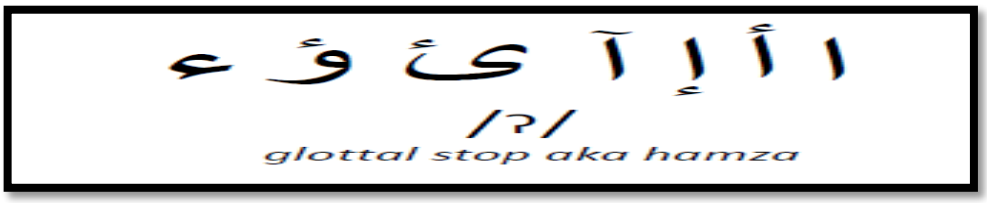


Figure 30: letters (form + mark) in non-Distinctive format. [Mona Diab & Nizar Habash, 2014]

- Arabic script uses a set of optional diacritics

Vowel				Nunation			Gemination
بَ	بُ	بِ	بْ	بً	بٌ	بٍ	بّ
/ba/	/bu/	/bi/	/b/	/ban/	/bun/	/bin/	/bb/

Table 1: A set of optional diacritics. [Mona Diab & Nizar Habash, 2014]

- Combinable
  - /kattab/ to dictate كَتَّب

- Putting it together:
  - Simple combination

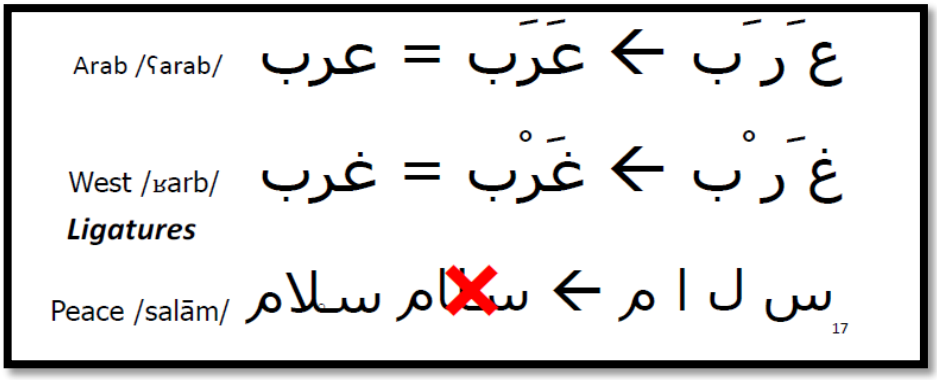


Figure 31: Combination example. [Mona Diab & Nizar Habash, 2014]

2. **Arabic Numerals:**
   - Decimal system
   - Numbers written left-to-right in right-to-left text



Figure 32: Numbers written format. [Mona Diab & Nizar Habash, 2014]

   - Three systems of enumeration symbols that vary by region

Western Arabic *Tunisia, Morocco, etc.*	0	1	2	3	4	5	6	7	8	9
Indo-Arabic *Middle East*	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
Eastern IndoArabic *Iran, Pakistan, etc.*	٠	١	٢	٣	۴	۵	۶	٧	٨	٩

Table 2: Type of systems enumeration symbols. [Mona Diab & Nizar Habash, 2014]

4. Works on Deep learning for Arabic NLP:

According to the writer "Mahmoud Al-Ayyoub Aya Nuseir Kholoud Alsmearat Yaser Jararweh Brij Gupta "from the article" Deep learning for Arabic NLP: A survey ", she cited many of the work done by other people in deep learning for Arabic NLP, each work with these references4. Deep learning for Arabic NLP:

In this section some of work of Arabic Natural Language Processing solved by deep learning will be presented, in each type one or two method (experiments) is put and the of solutions:

4.1. Optical Character Recognition (OCR):

“In this section, we discuss the works employing DL techniques for Optical Character Recognition (OCR) problems in ANLP.”

4.1.1. Handwritten Text:

"We start our discussion with the one of the most frequently addressed OCR problems, which handwritten Arabic text recognition. In one of earliest works, Al-Jawfi [Al-Jawfi, 2009] introduced an approach for Arabic handwritten text recognition problem. Based on the work of LeCun et al. [LeCun et al, 1990], Al-Jawfi's approach used LeNet, which is a convolutional neural network (CNN). The approach involves two phases: one for recognizing the body of the characters and one for recog-nizing the dots. It takes as input a set of 16 x 16 normalized images. They go through two hidden layers and the backpropagation algorithm was used to update the weights in the network. The experiments on this network was carried out using a dataset of 758 segmented images of Arabic characters (550 for training and 208 for testing) handwritten by different people. On average, the Mean Square Error (MSE) for the training set was 0.087 and, for the testing set, it was 0.42. The results are not very encouraging and the author do not compare them with the state-of-the-art results on the same dataset."

4.1.2. Printed Text:

"The authors of [Rashid et al, 2013] addressed the problem of recognizing multi-font printed Arabic text in low-resolution images. The experimented with three types of networks inspired by [Graves et al, 2007], [Graves et al, 2009], [Graves, 2012]: MDRNN, MDLSTM, and CTC. The dataset they used is taken from the Arabic Printed Text Image (APTI) dataset [Slimane et al, 2009] and it contained ten different fonts and six different font sizes. As for evaluation they used the tasks of the Arabic text recognition competition of ICDAR 2013[1]. The results showed that their system achieved recognition rates of 99% or above on most considered fonts and font sizes."

4.1.3. Video-Overlaid Text:

"The work of Yousfi et al. [Yousfi et al, 2014], [Yousfi et al, 2015], [Yousfi et al, 2015], [Yousfi, 2017] addressed the problem of detecting and recognition of Arabic text overlaid in videos. In [Yousfi et al, 2014] the authors proposed to use a CNN inspired by a prior work of the same authors [Delakis, 2008].They showed the superiority of the CNN approach compared with two approaches relying on multiexit asymmetric boosting cascade. The developed CNN consists of six layers with the first four layers being used for features extraction and the remaining two layers being used for classification. The authors used an in-house dataset collected from Arabic TV channels. The training set consisted of 30,000 text images and each method used bootstrapping to obtain negative examples.

The dataset had two test sets: one has 201 images with 959 annotated text areas while the other has 164 images with 1017 annotated text areas. For the first test set, the CNN approach gave better F-measure and lower computation time compared with the other two approaches. However, its detection rate (DR) was lower. As for the second test set, CNN results were better in terms of F-measure and DR."

[1] http://diuf.unifr.ch/diva/APTI/competitionICDAR2013.html

4.1.4. Natural Scene Text:

"The last work discussed in the previous subsection [Mohit et al, 2017] addressed the problem of detecting and recognizing text in natural scene images and mentioned the IIIT Arabic scene text dataset. In this section, we survey other papers addressing the same problem or problems relevant to it.

The authors of [Gomez et al, 2016], [Gomez et al, 2017] addressed the problem of detecting text in natural scene images, which can be more difficult than text printed on documents or embedded in videos due to the different levels of perspective distortion, illumination, etc. The problem's difficulty is increased when considering multiscript text on complex backgrounds. Following a patch based approach, in [Gomez et al, 2016], they proposed to combine single layer CNN with a Naive-Bayes Nearest Neighbor (NBNN) classifier. For experimentation, they introduced their own MLe2e dataset[2], which consists of 711 scene images covering four different scripts (not including Arabic). They also experimented with the Video Script Identification Competition (CVSI-2015) dataset[3] [Sharma et al, 2015], which consists of 10,000 pre-segmented video words equally distributed among the ten scripts it covers (including Arabic).

The experiments on both datasets showed the superiority of the proposed approach compared with several state-of-the-art approaches. However, for the CVSI-2015 dataset, the approach of the Google team produced slightly higher results."

4.2. Caption Generation:

"The problem of automatically generating a description for a given image is an interesting and practical one. Unfortunately, it is almost untouched in the ANLP community. In [Jindal et al, 2017], the author addressed this problem for Arabic by associating fragments of the image with root words. The fragments were extracted using a DNN pretrained on the ImageNet dataset [Deng et al, 2009].DBN were used to find the most suitable words to be associated with the image and the different root words to be associated with the fragments. Finally, the caption sentences are generating by using dependency tree relations. For evaluation, the author created two datasets. The first one was created by taking 10,000 images from the ImageNet dataset and have human experts add a caption to each image. As for the second dataset, it was created by taking 10,000 images from the Al-Jazeera news website (which has both Arabic and English content).

The results were very encouraging as they represent the first approach for Arabic caption generation. Moreover, they were much higher than the simple approach of generating English captions and automatically translating them into Arabic."

4.3. Automatic Speech Recognition (ASR):

"In one of the earliest works employing DL for ASR, the author of [Alotaibi et al, 2004] used RNN to recognize spoken Arabic digits. The RNN used was based on [Elman, 1990] and it consisted of two hidden layers. The proposed model performed well on an in-house dataset for both multi-speaker and speaker-independent settings.

[2] https://github.com/lluisgomez/script_identification
[3] http://www.ict.griffith.edu.au/cvsi2015/

The authors of [Kingsbury et al, 2012] describe a distributed NN training algorithm, based on Hessian-free optimization, that scales to deep networks and large datasets. For the state-level minimum Bayes risk (sMBR) criterion, the authors state that this training algorithm is faster than stochastic gradient descent by a factor of 5.5 and yields a 4.4% relative improvement in word error rate on a 50-hour broadcast news task. The authors also state that distributed Hessian-free sMBR training yields relative reductions in word error rate of 713% over cross-entropy training with stochastic gradient descent on two larger tasks: Switchboard and DARPA RATS noisy Levantine Arabic. Their best Switchboard DBN achieved a 16.4% WER."

4.4. Language Modeling:

"In previous sections, we discussed the use of LM in different problems (e.g., [Yousfi et al, 2017] used them for OCR and [Amr El-Desoky et al, 2013] used them for ASR). However, building LM can be studied independently. In [Kim et al, 2016], the authors proposed a character-level LM that can work on English as well as MRL such as Arabic. The proposed model applies CNN on input characters before feeding them into LSTM RNN-LM. The results for the Arabic language showed that the proposed LM outperformed various baselines working on word level or morpheme level. The authors have made the implementation of their approach publicly available[4]."

4.5. Automatic Machine Translation (ATM):

"In this section, we discuss the efforts invested in developing DL based ATM systems for translating text from/to the Arabic language. We have noticed a scarcity of publications on this subject despite the popularity of DL approaches in ATM systems (especially, the commercial ones such as Google's). There are other relevant problems that have yet to be addressed using DL such as the translation between Arabic dialects [Harrat et al, 2017].

The authors of [Deselaers et al, 2009] proposed to address the Arabic-English machine transliteration problem using DBN, which contains multiple layers of RBM. The proposed approach has three important parts. The first one is the source encoder, which deals with source words by converting them to dimensional binary vectors, then 460 feeding them into first layer in the source encoder, the output of each layer is considered as an input to the next layer. The second part called joint layer. This layer uses the output of the source encoder as an input in order to get a state of hidden neurons, and infer an output state to use as input to the top level of the output encoder. The third part is the target encoder. Within this part, the output vector is decoded by traversing down words through the output encoder. Because RBMs are bidirectional models, the translation process could be from Arabic to English, and vice versa. During the experiments, the authors studied the effects of network structure, by making the number of layers and the size of the bottom layers in the source and target encoders fixed, but the joint layer could be in different layers and different sizes."

[4] https://github.com/yoonkim/lstm-char-cnn

4.6. Text Categorization (TC):

"TC is a fundamental problem in information retrieval and NLP. However, not many works have proposed to use DL to address it. The only exceptions are [Jindal et al, 2016], [Elrazzaz et al, 2017]. The authors of [Jindal et al, 2016] developed an approach for Arabic text categorization that is based on three stages. The first stage is preprocessing, where the punctuation marks, auxiliary verbs, pronouns, etc., were removed. Then, the remaining words are represented in root patterns using a letter weight and order scheme.

The second stage is clustering. It is done into two phases using Fuzzy C-Means (FCM) clustering and Markov clustering. The third stage is training DBN for each cluster. They evaluated their approach using two datasets: Al-Jazeera dataset (consisting of 10,000 articles) and the Saudi Press Agency dataset (consisting of 6,000 articles). The author reported an F-measure of 91.02%, which is significantly higher than the results reported in the literature."

4.7. Sentiment Analysis (SA):

"In [Dahou et al, 2016], the authors focused on building a word embeddings using a 3.4 billion words corpus from a collected 10 billion words web-crawled corpus. They used a CNN trained on top of pretrained Arabic word embeddings for SA to evaluate the quality of these word embeddings. The experiment results show that their scheme outperforms existing methods on several publicly available datasets [Aly et al, 2013], [Nabil et al, 2015], [Refaee et al, 2014], [Abdulla et al, 2013], [ElSahar et al, 2015].

Relying on word embeddings as the main source of features for SA, the authors of [Altowayan et al, 2016] presented their system which consists of the following steps. First, they compile a large Arabic corpus from various sources to learn word representations.

Second, they train and generate word vectors (embeddings) from the corpus using the word2vec tool [Mikolov et al, 2013].Third, they use the embeddings in their feature representation for training several binary classifiers to detect subjectivity and sentiment in both Standard and Dialectal Arabic. The implementation of this approach is publicly available[5]. The authors compare their results with other methods in literature. Their approach, which does not employ hand-crafted features, achieved a slightly better accuracy than the top hand-crafted methods."

4.8. Question Answering (QA):

"In [Romeo et al, 2017], the authors addressed the question ranking problem in Arbic community QA (cQA) forums. They used the Farasa toolkit[6] to build a UIMA[7] based processing pipeline that involves a Tree-Kernel (TK) based ranker. Then, they used LSTM networks to choose the question fragments to be used in the ranker. They complement the features they compute with word embeddings features computed using the word2vec tool [Mikolov et al, 2013]. They evaluated their approach on the CQA-MD corpus, a medical Arabic corpus that is part of the SemEval 2016 Task 3-D [Nakov et al, 2017].The results show strong performance by the proposed pipeline, which is further enhanced by the use of LSTM networks."

[5] https://github.com/iamaziz/ar-embeddings
[6] http://farasa.qcri.org/
[7] https://uima.apache.org

4.9. Distributed Word Representations:

"The authors consider different techniques to build vectorized space representations for Arabic. Specifically, the authors built three models: continuous bag of word (CBOW), Skipgram (SKIP-G), and GloVe [Mikolov et al, 2013], [Mikolov et al, 2013], [Pennington et al, 2014]. They published these mode for public use[8]. The authors studied the impact of these representations on two NLP tasks: Query Expansion for Information Retrieval and Short Answer Grading, and the results were encouraging." [Mahmoud Al-Ayyoub et al, 2017]

Type	Problem	Author	Article name	Date	Works	Results
Optical Character Recognition (OCR).	Handwritten Arabic text recognition.	Al-Jawfi.	"Handwriting arabic character recognition lenet using neural network."	2009	Introduced an approach for Arabic handwritten text recognition problem. Based on the work of LeCun et al, 1990: "Handwritten digit recognition with a back7 propagation network, Al-Jawfi's approach used LeNet, which is a convolutional neural network (CNN). The approach involves two phases: one for recognizing the body of the characters and one for recog-nizing the dots.	The results are not very encouraging and the author does not compare them with the state-of-the-art results on the same dataset.

[8] https://sites.google.com/site/mohazahran/data

The results showed that their system achieved recognition rates of 99% or above on most considered fonts and font sizes.	CNN results were better in terms of F-measure and DR.
The experimented with three types of networks inspired by A. Graves et al, 2007: Offline handwriting recognition with multidimensional recurrent neural networks, A. Graves et al, 2009: Offline handwriting recognition with multidimensional recurrent neural networks, A. Graves et al, 2012: Offline arabic handwriting recognition with multidimensional recurrent neural networks: MDRNN, MDLSTM, and CTC. As for evaluation they used the tasks of the Arabic text recognition competition of ICDAR 2013.	The authors proposed to use a CNN inspired by a prior work of the same authors [M. Delakis, 2008: text detection with convolutional neural networks.
2013	2017
“Low resolution arabic recognition with multidimensional recurrent neural networks.”	“Arabic text detection in videos using neural and boosting-based approaches: Application to video indexing.”
S. F. Rashid, M.-P. Schambach, J. Rottland, S. von der Null.	S. Yousfi, S.-A. Berrani, C. Garcia.
Recognizing multi-font printed Arabic text in low-resolution images.	Addressed the problem of detecting and recognition of Arabic text overlaid in videos. (Video-Overlaid Text).

Showed the superiority of the proposed approach compared with several state-of-the-art approaches. However, for the CVSI-2015 dataset, the approach of the Google team produced slightly higher results.	The results were very encouraging as they represent the first approach for Arabic caption generation. Moreover, they were much higher than the simple approach of generating English captions and automatically translating them into Arabic.
For experimentation, they introduced their own MLe2e dataset, which consists of 711 scene images covering four different scripts (not including Arabic). They also experimented with the Video Script Identification Competition (CVSI-2015) dataset [N. Sharma et al, 2015], which consists of 10,000 pre-segmented video words equally distributed among the ten scripts it covers (including Arabic).	The author addressed this problem for Arabic by associating fragments of the image with root words. The fragments were extracted using a DNN pretrained on the ImageNet dataset [J. Deng et al, 2009]: Imagenet: A large-scale hierarchical image database.DBN were used to find the most suitable words to be associated with the image and the different root words to be associated with the fragments. Finally, the caption sentences are generating by using dependency tree relations.
2017	2017
"Unconstrained scene text and video text recognition for arabic script."	"A deep learning approach for arabic caption generation using roots-words."
M. Jain, M. Mathew, C. Jawahar.	V. Jindal.
Detecting and recognizing text in natural scene images and mentioned the IIIT Arabic scene text dataset.	The problem of automatically generating a description for a given image is an interesting and practical one. Unfortunately, it is almost untouched in the ANLP community.
	Caption Generation.

Yields a 4.4% relative improvement in word error rate on a 50-hour broadcast news task. The authors also state that distributed Hessian-free sMBR training yields relative reductions in word error rate of 713% over cross-entropy training with stochastic gradient descent.	The results for the Arabic language showed that the proposed LM outperformed various baselines working on word level or morpheme level. The authors have made the implementation of their approach publicly available.
Used RNN to recognize spoken Arabic digits. The RNN used was based on [J. L. Elman, 1990]: Finding structure in time and it consisted of two hidden layers. The proposed model performed well on an in-house dataset for both multi-speaker and speaker-independent settings.	The authors proposed a character-level LM that can work on English as well as MRL such as Arabic. The proposed model applies CNN on input characters before feeding them into LSTM RNN-LM.
2004	2016
Spoken arabic digits recognizer using recurrent neural networks, in: Signal Processing and Information Technology.	“Character-aware neural languagemodels.”
Y.A.Alotaibi.	Y. Kim, Y. Jernite, D. Sontag, A. M. Rush.
Employing DL for ASR.	Building LM can be studied independently.
Automatic Speech Recognition (ASR).	Language Modeling.

/	The author reported an F-measure of 91.02%, which is significantly higher than the results reported in the literature.
Proposed to address the Arabic-English machine transliteration problem using DBN, which contains multiple layers of RBM. The proposed approach has three important parts. The first one is the source encoder, The second part called joint layer. The third part is the target encoder. The authors studied the effects of network structure, by making the number of layers and the size of the bottom layers in the source and target encoders fixed, but the joint layer could be in different layers and different sizes.	Developed an approach for Arabic text categorization that is based on three stages. The first stage is preprocessing, where the punctuation marks, auxiliary verbs, pronouns, etc., The second stage is clustering. It is done into two phases using Fuzzy C-Means (FCM) clustering and Markov clustering. The third stage is training DBN for each cluster. 6,000 articles).
2009	2016
“A deep learning approach to machine transliteration.”	“A personalized markov clustering and deep learning approach for arabic text categorization.”
T. Deselaers, S. Hasan, O. Bender, H. Ney.	V. Jindal.
Address the Arabic-English machine transliteration problem using DBN.	TC is a fundamental problem in information retrieval and NLP.
Automatic Machine Translation (ATM).	Text Categorization (TC).

The experiment results show that their scheme outperforms existing methods on several publicly available.
The authors focused on building a word embeddings using a 3.4 billion words corpus from a collected 10 billion words web-crawled corpus. They used a CNN trained on top of pretrained Arabic word embeddings for SA to evaluate the quality of these word embeddings. Second, they train and generate word vectors (embeddings) from the corpus using the word2vec tool [T. Mikolov et al, 2013]. , Linguistic regularities in continuous space word representations. Third, they use the embeddings in their feature representation for training several binary classifiers to detect subjectivity and sentiment in both Standard and Dialectal Arabic.
2016
“Word embeddings and convolutional neural network for arabic sentiment classification.”
A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud, P. Duan.
Sentiment Analysis (SA).
Sentiment Analysis (SA).

The results show strong performance by the proposed pipeline, which is further enhanced by the use of LSTM networks.	The results were encouraging.
They used the Farasa toolkit to build a UIMA based processing.	The authors consider different techniques to build vectorized space representations for Arabic. Specifically, the authors built three models: continuous bag of word (CBOW), Skipgram (SKIP-G), and GloVe They published these mode for public use. The authors studied the impact of these representations on two NLP tasks: Query Expansion for Information Retrieval and Short Answer Grading.
“Language pro1cessing and learning models for community question answering in arabic, Information Processing & Management.”	/
S. Romeo, G. Da San Martino, Y. Belinkov, A. Barron-Cedeno, M. Eldesouki, K. Darwish, H. Mubarak, J. Glass, A. Moschitti.	2013/2014
Addressed the question ranking problem in Arbic community QA (cQA) forums.	Build vectorized space representations for Arabic.
Question Answering (QA).	Distributed Word Representations.

+“Distributed representations of words and phrases and their compositionality.”

+“Efficient estimation of word representations in vector space.”

+“Manning, Glove: Global vectors for word representation.”

T. Mikolov, K. Chen, G. Corrado, J. Dean, I. Sutskever, J. Pennington, R. Socher, C.Manning,S.wen-tan Yih, Z.geoffrey.

Table 3: Works on Deep learning for Arabic NLP Summarization.

5. Related work:

We present in this section, some research work about the improvement of the answer question as well as work on question answering and word2vec or both together.

[Lahsen Abouenour, Karim Bouzoubaa, Paolo Rosso. December 2008]

Their approach focus on the improvement of Q/A using Arabic WordNet, the adoption of the Amine AWN ontology traces new ways to get the answer expected by exploiting the

definition of a concept, its synonyms and its context represented by the hierarchical nearly types.

To make the approach, they make an experiment, so they take a set of 82 questions of CLEF that was translated into Arabic, classified into different domains (sport, geography, politic, etc.).

After producing a set of new keywords based on Amine AWN, the purpose of AWN is approach expend the query of the user not only at the morphological level but at the semantic one also.

They look for the correct answer in the first fixe snippets returned by google, and then to evaluate the results they use two measures:

1) Accurancy which is the average of questions where we find the right answer in the first snippet.
2) The mean Reciprocal Rank (MRR): the reciprocal rank of query response is the multiplicative inverse of the rank of the correct answer and MRR is the average of the reciprocal rank of results for a sample of queries.

The results shows for 19,5%of questions success to get the expected answer in the first five snippets , to get the expected answer in the first 5 snippets using their semantic expansion after failing to get it without any QE.

[Yonatan Belinkov, Mitra Mohtarami, Scott Cyphers, James Glass, 2015]

They represented words, phrases, sentences and whole questions and answers in vector space by using word2vec, and computed various features from them for a classifier, for both English and Arabic. They showed the utility of these vector-based features for addressing the answer selection and the YES/NO answer inference tasks in community question answering.

[Hani Al-Chalabi, Santosh Ray, Khaled Shaalan, 2015]

They are presenting a method to add semantically equivalent keywords in the questions by using semantic resources. The experiments suggest that the proposed research can deliver highly accurate answers for Arabic questions.

This work can be extended to improve the AWN and study the applicability of improved version of AWN. They focused only on designing and developing Question Analysis module of Arabic Question Answering Systems.

[Marwan Torki, Maram Hasanain, Tamer Elsayed, 2017]

The approach describes the system that they developed to participate in SemEval-2017 Task 3 on Community Question Answering. Their system has focused on the Arabic Subtask D which is confined to Answer Selection in Community Question Answering, i.e., finding good answers for a given new question.

They have adopted a supervised learning approach where linear SVM models were trained over similarity features. In their best submission, term-based similarity features and word2vec similarity features were both used; their system ranked second among the other participating teams.

6. Conclusion:

In this chapter, Deep learning for Arabic natural language processing as a general idea was presented. So in this chapter on: natural language processing, Arabic natural language processing, deep learning for Arabic natural (work) language processing is discussed.

Then, for natural language processing: definition, applications, examples, and the stages of natural language processing were tackled. In addition, for Arabic natural language processing: definition, grammar, type. After that, the purpose of this chapter, which is to know the use of deep learning in Arabic natural language in different domains was discussed too, and with some work related to our title repport is ended.

Chapter IV: Proposed Approach

1. Introduction:

After the creation of word2vec, it is used in several domains, in approach below also we use word2vec with other techniques to reformulate the requests and to see if there is a good improvement or not, because this technique is new, for that in this chapter we detail the general approach structure proposed, and explain each step in detail.

2. Proposed approach:

First, in this step, we present the different phases followed to design our proposed approach, furthermore explaining the details of each phase, starting by giving a general diagram on our approach.

The functional implementation of the approach relies on several modules for the application of the Word2vec algorithm to improve queries of Arabic Question/Answering system.

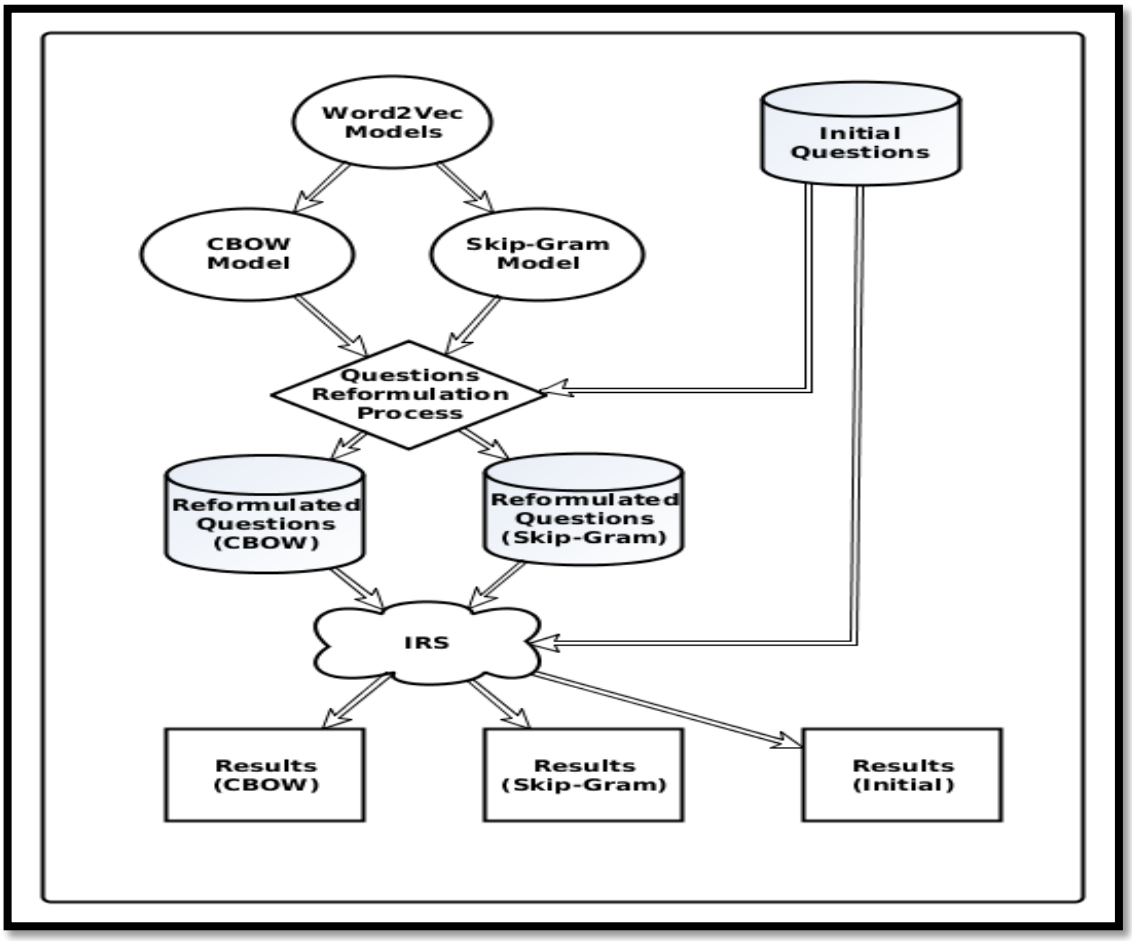


Figure 33: General architecture of our proposed approach.

Our approach is the definition of a method of improvement of the questions in Arabic language, by the use of word2vec like a main tool. And after that we do a test to see if we reach our goal or not.

To start the approach, in the first place takes the questions without reformulation, after that try to find answers for each question by using IRS. For each question takes the first N links which contain the answers.

In second place, takes the questions another time and make for them a few changes (reformulation), this changes based on using two word2vec models (CBOW and Skip Gram) each time a models, after try to find answers for each question, and take the first N links which contain the answers too.

In the end, do a test to see if there is an improvement or not between the origin and reformulated questions.

2.1. Dataset:

2.1.1. Dataset Definition:

A data set is a set of data that is collected for a specific purpose. There are many ways in which data can be collected: for example, as part of service delivery, one-off surveys, interviews, observations, and so on. In order to ensure that the meaning of data in the data set is clearly understood and data can be consistently collected and used. [Australian Institute of Health and Welfare (AIHW), 2007].

So, we work on a collection which contains a number of questions in different field with Arabic language.

2.2. Questions:

2.2.1. Question Definition:

A question is a linguistic expression used to make a request for information, or the request made using such an expression. The information requested is provided in the form of an answer. [17]

2.2.2. Question Type:

- **Factoid queries**: WH questions like when, who, where.
- **Yes/ No queries**: Is Berlin capital of Germany?
- **Definition queries**: what is leukemia?
- **Cause/consequence queries**: How, Why, What. What are the consequences of the Iraq war?
- **Procedural queries**: which are the steps for getting a Master degree?
- **Comparative queries**: what are the differences between the model A and B?
- **Queries with examples**: list of hard disks similar to hard disk X.
- **Queries about opinion**: What is the opinion of the majority of Americans about the Iraq war? [Shekarpour, 2015]
- **Indirect Requests** (*I would like you to list …*)
- **Commands** (*Name all the presidents …*)
- **Summaries** (*What are the arguments for and against…*) [Magnini, 2005]

2.2.3. Question part:

The questions in our database consist of 4 parts: question words, doubted words, central words, prepherical words.

- **Question words:** These are the question tools that come on the first part of the question(WH Question), for examples: اين, متى,كم
- **Stop words:** its contains several types:
  - Reference names like: هذه, هذا
  - Related names like: الذي
  - Prepositions like: من, إلى, على
- **Central words:** Are the key words in the questions, and they are taken apart, and by deleting them, the meaning is lost while the answer contains a relationship with those words, such as the named entity.
- **Prepherical words:** are the characteristic of question when we suppresses, the meaning don't change.

And we will represent this part in the example below:

كم ينفق على التدخين سنويا في الولايات المتحدة الامريكية؟

How much is spent on smoking annually in the United States?

Words	Type
كم	Question words
ينفق	Central words
على	Stop words
التدخين	Central words
سنويا	Prepherical words
في	Stop words
الولايات المتحدة الامريكية	Central words

Table 4: Presentation of the question example type.

2.3. Unique encoding of texts:

The unique encoding of texts in standard format makes it possible to represent the texts without any deformation at the level of character during reading.

All the texts of our corpus are represented with a UTF-8[9] encoding.

Indeed, in this part**,** we explain our test detail in general, but also we take each step and talk about it.

2.4. Questions reformulation process:

This process is using to reformulate the questions which is the output of work, it consist of 11 steps:

1. Normalization.
2. Skip words between quotation marks.
3. Remove "Stop words".
4. Skip Named Entities.
5. Put Named Entities between quotation marks.
6. Transformation number to string (Int to Str).
7. Transformation string to number (Str to Int).
8. Extract Similar Words to the rest words.

[9] The Unicode Transformation Format is also known as the Universal Transformation Format.

9. Stemming for similar words.
10. Determination the threshold of Similarity.
11. Use of “+”sign or “ ” (Space) sign.

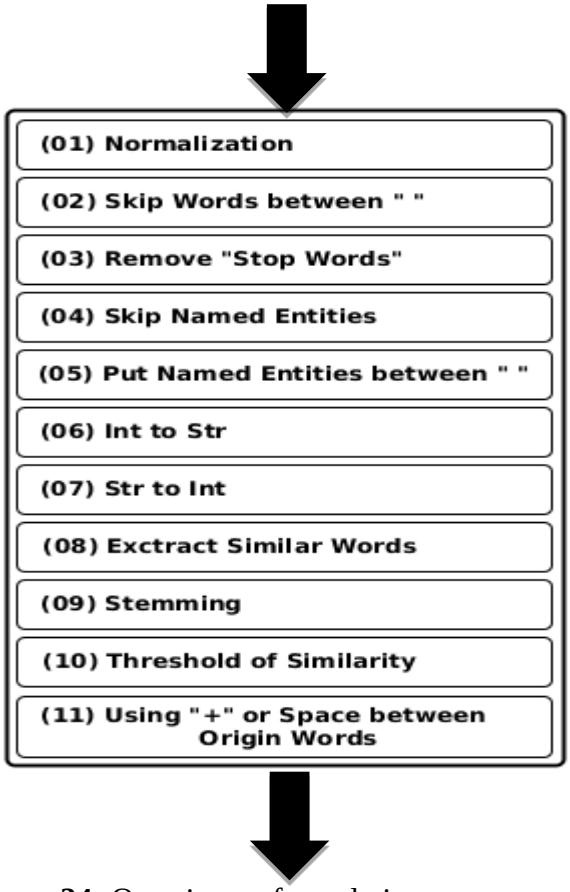


Figure 34: Questions reformulation process steps.

2.4.1. Normalization:

Normalization or normalisation refers to a process that makes something more normal or regular. [18]

Furthermore, Normalization is the process by which you can perform certain transformations of text to make it reconcilable in a way which it may not have been before.

Normalization of Arabic characters is a common preprocessing step when dealing with Arabic text.

Here are some of normalization examples:

- Letters “أ”, “إ”, “آ”, transformed to “ا”.
- Letter “ة” transformed to “ه”.
- Letter “ى”transformed to “ي”.
- Remove the special characters, for example “-”transformed to “”, etc.

2.4.2. Skip words between quotation marks:

All the words into quotation marks, don't make for them extraction of similar words. Because, the words between quotation marks when put them in a search engine, so the search engine looks on an exact match for these words [19]. So, can't extract similar words for them.

2.4.3. Remove "Stop words":

Stop Words are words which do not contain important significance to be used in Search Queries. Usually these words are filtered out from search queries because they return vast amount of unnecessary information. [20]

Commonly used words that are excluded from searches to help index and parse web pages faster. Some examples of stop words are: "a", "and", "but", "how", "or", and "what." While the majority of all Internet search engines utilize stop words, they do not prevent a user from using them, but they are ignored.

For example, if you were to search for: "What is a motherboard?" on Computer Hope, the search engine would only look for the term "motherboard." [21]

Here are some examples:

- Question Tools like: How long كم, متى when, whereفي أين
- The names of the signal like: This هذه, This هذا
- Linked Names like:الذي Which
- Prepositions like:من From, to إلى , onعلى

In this step, the extraction of similar words for Stop Words give bad result, so remove of Stop words to see if can get good results.

2.4.4. Skip Named entities:

2.4.4.1. Named entity Definition:

In information extraction, a named entity is a real-world object, such as persons, locations, organizations, products, etc., that can be denoted with a proper name. It can be abstract or have a physical existence. Examples of named entities include Barack Obama, New York City, Volkswagen Golf, or anything else that can be named. Named entities can simply be viewed as entity instances (e.g., New York City is an instance of a city). As an example, consider the sentence, "Obama is the president of the United States". Both "Obama" and the "United States" are named entities since they refer to specific objects (Barack Obama and United States). However, "president" is not a named entity since it can be used to refer to many different objects in different worlds (in different presidential periods referring to different persons, or even in different countries or organizations referring to different people). [22]

According to the test of named entities when using word2vec, the named entities are fixed because they are central words in the questions(see Question part in chapter4), when apply the extraction of similar words on these words it gives other words different in meaning than the original so, the similarity in this case lead to a bad results.

However, the answers are based on the (central words) the words origins so, the fixation of these words gives good and a precise results.

Instead of extract similar words to named entities, leave them in the original case, for example of Named entity:الامارات UAE,الكبيسي Al Kubaisi,يونيسيف UNICEF, Stockholmستوكهولم , النهيانNahyan, etc.

2.4.5. Put Named Entity between quotation marks:

Every Named Entity exist we make them between quotation marks, because we have two types of Named Entities (simple and composed), the simple Named Entity is easy to treat but when we come to the composed Named Entity after search for question contain a composed Named Entity it divide to two words, so all the meaning of question change, for that every Named Entities we put them into quotation marks.

2.4.6. Transformation number to string (Int to Str):

Every number in the questions transform to string, For example:

23 → (23 | ثلاث و عشرون)

1985 → (1985 | ألف و تسعمئة و خمس و ثمانون)

2.4.7. Transformation string to number (Str to Int):

On the contrary, every number in the questions is writing in letters (circumstance) transform to roman number, except the by using of roman numbers and leave number and the dates in the original case, For example:

الثالث→ III

2.4.8. Extract similar words:

Extract similar words, by the use of two models of Word2Vec, and then calculate the similarity by the cosine function.

For example the result of similar word for the word: "العوائد Returns" .and the figure (35) show the results:

CBOW Model	
Similarity	Word
0.747137188911	والارباح
0.716347932816	والعوائد
0.7038859725	الارباح
0.696322441101	والنفقات
0.687442064285	العائدات
0.684268593788	والمدفوعات
0.682197570801	وارباح
0.68206435442	والديون
0.680844783783	الفوائض
0.679539680481	والاقتراض

Skip-Gram Model	
Similarity	Word
0.828226327896	العائدات
0.806858599186	الارباح
0.801460146904	النفقات
0.772960424423	السيوله
0.76902794838	الايرادات
0.761806249619	المخصصات
0.756675601006	التكاليف
0.748052477837	القروض
0.739977121353	المداخيل
0.73487919569	المكاسب

Figure 35: Similar words of "العوائد" extracting by CBOW and Skip-Gram models of Word2Vec.

And here the algorithm of extraction similar words:

Input: A user query (Q), Word2Vec_Model (WM)
Output: Enhanced query (QE)
Step 1: Extract the Prepherical words from keywords Words W1, W2, ..., Wn of the user query Q.
Step 2: For i= 1 to n
Use Word2vec to extract top n semantically

equivalent terms for Prepherical Words consideration. For Prepherical Words Wi, semantically equivalent words are Ci1, Ci2, ..., Cim.

Step 3: Construct a new Query using Boolean operators "AND" and "OR" as (W11 | W12 |... | W1m) + (W21 | W22 |... | W2m) + ... + (Wn1 |Wn2 |... | Cnm).

Step 4: End

2.5. Two models of Word2Vec:

Word2Vec is a machine learning algorithm created by Google in 2013 [8], the purpose of Word2Vec is to group words that semantically similar in vector space.

Word2Vec model can create a very rich distributed representation of words that also preserves the semantic relationships that words have with other words. A train Word2Vec model can basically understand the meaning of a word based on the past appearances. The most famous example with be \(v(king) – v(man) + v(women) ≈ v(queen)\). This means that we can now cluster words, find similar words, and find similar relationships of different words and more. [6]

To know more details about Word2Vec and his calculations (see chapter I and II).
The two model of Word2Vec are:

2.5.1. Continuous bag-of-words (CBOW):

In the continuous bag-of-words architecture, the model predicts the current word from a window of surrounding context words. The order of context words does not influence prediction (bag-of-words assumption).

2.5.2. Skip gram:

In the continuous skip-gram architecture, the model uses the current word to predict the surrounding window of context words. The skip-gram architecture weighs nearby context words more heavily than more distant context words

According to the authors' note, CBOW is faster while skip-gram is slower but does a better job for infrequent words. [23]

After the extraction of similar words, the results contain: the "|" sign, and the "+" sign, in addition the "" sign.

The "|" sign is use to indicate the emplacement of "or" operator between similar words, its use to combine between similar words. The "|" sign combine between similar words, when put the "|" in a search engine, so the search engine combine searches as is planned in the" site 19"[19]

Then, the "" sign and the "+" sign use to indicate the emplacement of "and" operator is default concatenation operator, between the word or between similar with new word.

2.6. Similarity:

The model knows when two words are similar by the uses of cosine similarity to compare vectors.

After turning word into vectors, the cosine of the angle between word vectors can be used to compare how similar words are. The smaller an angle is, the larger (closer to 1) the cosine value will be. In this case, a larger cosine value indicates word vectors which have similar meanings.

When the angle between words is small, so the cosine of the angle between them is large, this means they are similar.

When the angle between them is large, so the cosine of the angle between them is small which means these words are not similar. [8]

2.4.9. Stemming (similar words):

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming is important in natural language understanding (NLU) and natural language processing (NLP). [24]Algorithms for stemming have been studied in computer science since the 1960s. [25]

In this step stemming is using to remove some words as example prefixes, for example: "والارباحProfits" → "ارباحProfits"

2.4.10. Determination the threshold of similarity:

As we said before the question is divided into 4 phases which is: WH Question, Central Words, Prepherical Words, Stop words. Our work is based on two phases Central words which is also divided into two phases Named Entitles and other words, and Prepherical words then give two values for each word for them. After extracting of similar words with its similarity, first, by playing with the values of the interval which contain two value one for the Prepherical similarity and the second for the central words similarity, for example: takes similar words that have similarity larger than or equal to 0.7 with 0.6, this gives the words closest to the origin word, also the number of similar words decreases.

These values can be changed several times, once 0.7 and other times 0.5, etc.

2.4.11. Use of "+"sign or " " (Space) sign:

As we said before, the "|" sign is use to indicate the emplacement of "or" operator between similar words, its use to combine between similar words. The "|" sign combine between similar words, when put the "|" in a search engine, so the search engine combine searches as is planned in the site [19]

Then, the "" sign and the "+" sign use to indicate the emplacement of "and" operator is default concatenation operator, between the word or between similar with new word.

After the extraction of similar words, for the first time we use the "|" sign and the "+" sign to extract all similar words.

Then after that, we change sometimes we use the "+" sign, or the "" sign to see if this give a good result.

2.7. Get answers from IRS:

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing. Information retrieval is the science of searching

for information in a document, searching for documents themselves, and also searching for metadata that describe data, and for databases of texts, images or sounds.
Automated information retrieval systems are used to reduce what has been called information overload. An IR systems is a software that provide access to books, journals and other documents, stores them and manages the document. Web search engines are the most visible IR applications.

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy. [26]

First, in this step try to get answers to each question, this step is done for the origin questions and for the reformulated questions with two models (CBOW and Skip Gram).

After each search done, store the N first answers given, for the three types (origin questions, reformulated questions(CBOW),reformulated questions(Skip Gram)),this work is done for the first 50 questions, and the figure below show the operation:

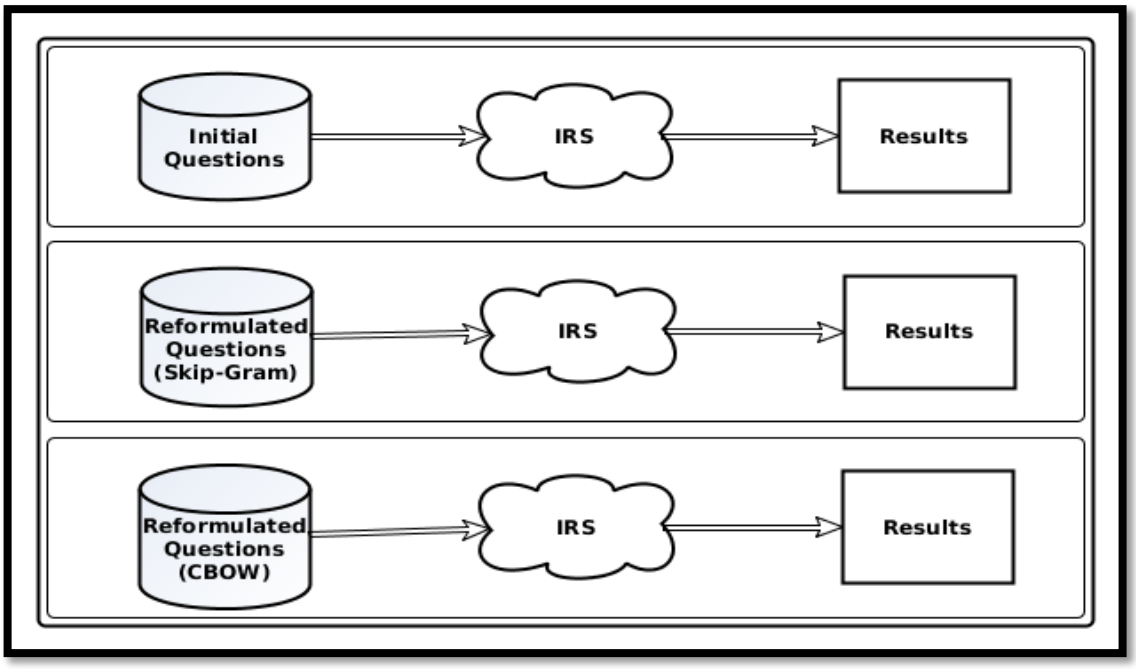


Figure 36: Extraction Architecture the results from IRS.

2.8. Testing the results:

To test the results get, two tests have been done.

2.8.1. Test1 (relevant/irrelevant):

The first test is a manual test (by expert) to extract relevant and irrelevant links which contain answers to the questions retrieved from IRS, for the initial questions (without reformulation), as well as the reformulated questions by the two models of Word2Vec (CBOW and Skip gram).and the (figure 37) show the work:

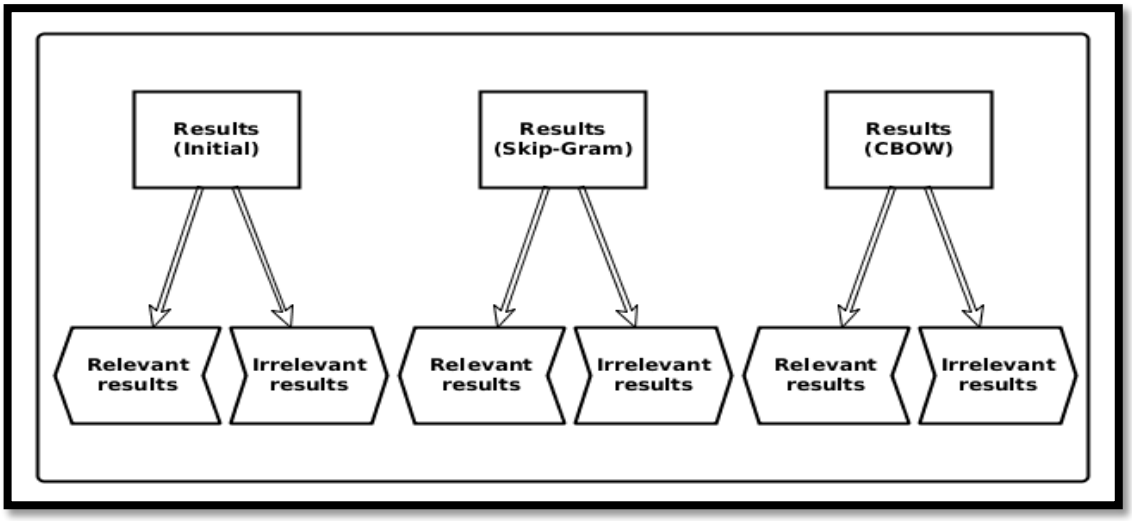


Figure 37: Manual test to extract relevant and irrelevant links.

2.8.2. Test2 (improvement/not improvement):

To measure our results in relation to initial query, we calculate:

- For each query we retrieve the number of relevance.
- In the second phase, we calculate the margin (difference) between the number of relevance for each initial question and that corresponds in query reformulated.
- we group the margins in different lists, a list for the questions that is have a result equal to 0, other equal to 1, then equal to 2(bigram), etc.
- After the extraction of the previous lists we calculate the avreage of each list.

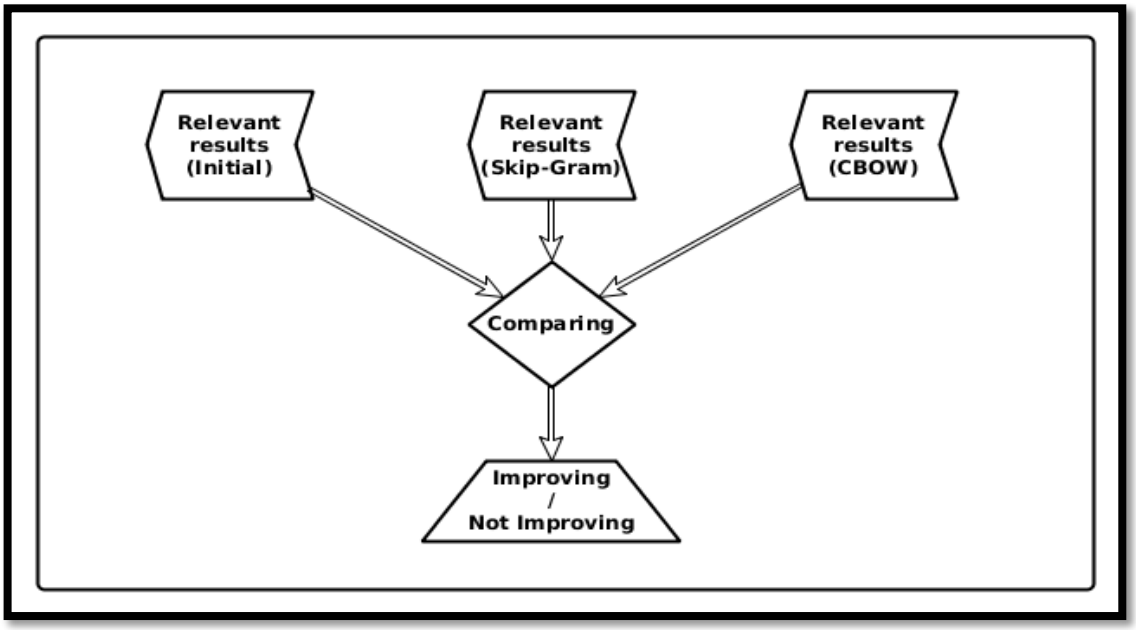


Figure 38: Architecture of the Comparison between the results to see if there is improvement.

3. Conclusion:

In this chapter, our proposed system has been described, which aims to design a system capable of improving Arab questions, our given approach, is expressed on a set of question in Arabic (the base).

This design will be put into operation in the following chapter.

Chapter V: Implementation and Evaluation

1. Introduction:

We present in this chapter the implementation of our proposed approach, we start with the presentation of the development environment, detail the different tools used, then we explain the steps of our implementation, in the end we interpret and comment the results obtained.

2. Development environment:

We present in this section, the programming languages used, also the development environment:

2.1. Python:

Python is a high-level, interpreted and general-purpose dynamic programming language that focuses on code readability. The syntax in Python helps the programmers to do coding in fewer steps as compared to Java or C++. The language founded in the year 1991 by the developer Guido Van Rossum has the programming easy and fun to do. The Python is widely used in bigger organizations because of its multiple programming paradigms. They usually involve imperative and object-oriented functional programming. It has a comprehensive and large standard library that has automatic memory management and dynamic features. [27]

2.1.1. Python Features:

2.1.1.1. Open source:

Python is publicly available open source software, anyone can use source code that doesn't cost anything.

2.1.1.2. Easy-to-learn:

Popular (scripting/extension) language, clear and easy syntax, no type declarations, automatic memory management, high-level data types and operations, design to read (more English like syntax) and write (shorter code compared to C, C++, and Java) fast.

2.1.1.3. High-level Language:

High-level language (closer to human) refers to the higher level of concept from machine language (for example assembly languages). Python is an example of a high-level language like C, C++, Perl, and Java with low-level optimization.

2.1.1.4. Portable:

High level languages are portable, which means they are able to run across all major hardware and software platforms with few or no change in source code. Python is portable and can be used on Linux, Windows, Macintosh, Solaris, FreeBSD, OS/2, Amiga, AROS, AS/400 and many more.

2.1.1.5. Object-Oriented:

Python is a full-featured object-oriented programming language, with features such as classes, inheritance, objects, and overloading.

2.1.1.6. Python is Interactive:

Python has an interactive console where you get a Python prompt (command line) and

interact with the interpreter directly to write and test your programs. This is useful for mathematical programming.

2.1.1.7. Interpreted:

A Python program are interpreted, takes source code as input, and then compiles (to portable byte-code) each statement and executes it immediately. No need to compiling or linking.

2.1.1.8. Extendable:

Python is often referred to as a "glue" language, meaning that it is capable to work in mixed-language environment. The Python interpreter is easily extended and can add a new built-in function or modules written in C/C++/Java code.

2.1.1.9. Libraries:

Databases, web services, networking, numerical packages, graphical user interfaces, 3D graphics, others.

2.1.1.10. Supports:

Support from online Python community [28]

2.2. Python module used:

2.2.1. Gensim Module:

Gensim is an open source Python library for Natural Language Processing. It provides an easy to load functions for pre-trained embeddings in a few formats and support of querying and creating embeddings on a custom corpus. Before feeding the raw data to your training algorithm, you might want to do some basic preprocessing on the text. [29]

We used “genism” to work with all function related to word2vec algorithms.

2.2.2. Re Module:

Regular expressions are basically just a sequence of characters that can be used to define a search pattern for finding text. This "search engine" is embedded within the Python programming language (and many other languages as well) and made available through the re module. [30]

Regular expressions are a powerful language for matching text patterns. The Python "re" module provides regular expression support. [31]

We used “re” to remove longation.

2.2.3. Unicodedata Module:

Unicode Character Database (UCD) is defined by Unicode Standard Annex #44 which defines the character properties for all unicode characters. This module provides access to UCD and uses the same symbols and names as defined by the Unicode Character Database.

- unicodedata.normalize (form, unistr):
  This function returns the normal form for the Unicode string unistr. Valid values for form are 'NFC', 'NFKC', 'NFD', and 'NFKD'. [32]

Example: (we use to set the ASCII code)

```
from unicodedata import normalize
print '%r' % normalize('NFKD', u'\u2460')
```
Output:
u'1'

The normal form KD (NFKD) will apply the compatibility decomposition, i.e. replace all compatibility characters with their equivalents. [33]

We used it to normalizing by NFKD format to Benajiba dataset.

2.2.4. Pyarabic Module:

A specific Arabic language library for Python provides basic functions to manipulate Arabic letters and text, like detecting Arabic letters, Arabic letters groups and characteristics, remove diacritics etc.

2.2.4.1. Features:

- Text tokenization
- Strip Harakat (all, except Shadda, tatweel, last_haraka)
- Sperate and join Letters and Harakat
- Reduce tashkeel
- Mesure tashkeel similarity (Harakats, fully or partially vocalized, similarity with a template)
- Letters normalization (Ligatures and Hamza)
- Numbers to words
- Extract numerical phrases
- Pre-vocalization of numerical phrases
- Unshiping texts [34]

We used it for working with Arabic words and numbers.

2.2.5. Python string Module:

The string module contains a number of useful constants and classes, as well as some deprecated legacy functions that are also available as methods on strings. [site35, 2018]

We used "string Module" to remove punctuation.

2.2.6. Google APIs Module:

Google APIs is a set of application programming interfaces (APIs) developed by Google which allow communication with Google Services and their integration to other services. Examples of these include Search, Gmail, Translate or Google Maps. Third-party

apps can use these APIs to take advantage of or extend the functionality of the existing services. [36]

Google web APIs enables developer using Google as a resource in their application to easily find and manipulate information on the web. [37]

We used google-api for get answers from google search-engine.

2.2.7. Matplotlib Module:

Matplotlib is a Python 2D plotting library which produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell (à la MATLAB or Mathematica), web application servers, and various graphical user interface toolkits. [38] We used it for show graphs.

2.2.8. Tahsphyne Module:

Tahsphyne is a finite state automaton stemmed based, it extracts affixes (prefixes and suffixes), with a predefined affixes list. It can all be done from a word and quote all possible configuration.

2.2.8.1. Features:

- Arabic word Light Stemming.
- Root Extraction.
- Word Segmentation
- Word normalization
- Default Arabic Affixes list.
- A customizable Light stemmer: possibility of change stemmer options and data.
- Data independent stemmer. [39]

We used “Tahsphyne” to do the stemming.

2.2.9. PANDA Module:

PANDA is an open-source Platform for Architecture-Neutral Dynamic Analysis. It is built upon the QEMU whole system emulator, and so analyses have access to all code executing in the guest and all data. PANDA adds the ability to record and replay executions, enabling iterative, deep, whole system analyses. [40]

We used PANDA to organize the Data frame.

2.2.10. Numpy Module:

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities [41]

We used “Numpy” to use vectors of Numpy.

2.2.11. Scikit-learn:

Scikit-learn is a Python module for machine learning built on top of SciPy and distributed under the 3-Clause BSD license. [42]

We use two functions from this module:

PCA: algorithm for representation of data multi-dimensional 2D.

TSNE: Package for the graphical.

3. Explanation of our approach:

3.1. Benajiba Dataset:

We work on a collection[10] created by Benajiba[11] which contains a number of questions in different field with Arabic language.

This Dataset is chosen because there's many works on this dataset.his words are simple, easy.

3.2. Benajiba questions:

Benajiba Dataset contains in total 193 questions, with different type: 191WH Questions and 2 Commands Questions. These questions are in different field, each questions has 3 words in minimum and 16 words in maximum.

We work with the first fifty questions from the Benajiba Dataset.

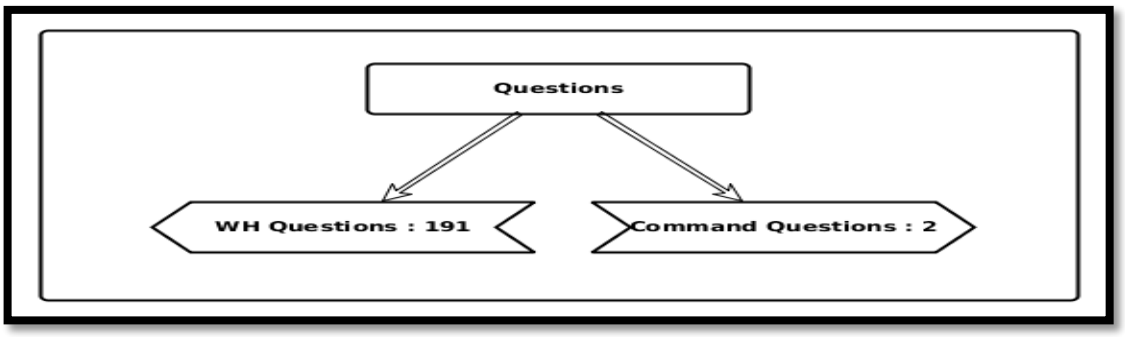


Figure 39: Types of questions use.

And here are some of these questions:

Number	Questions	Traduction
D 0005 AR AR	كم يبلغ ارتفاع برج ايفل؟	How high is the Eiffel Tower?
D 0116 AR AR	ما هو عنوان اول قصة قصيرة نشرت للكاتب محمد مستجاب؟	What is the title of the first short story published by writer Mohammed Mustap?
D 0012 AR AR	كم تبلغ قوة انفجار قنبلة نووية؟	What is the power of a nuclear bomb?
D 0014 AR AR	كم عدد الاعضاء الغير الدائمة في مجلس الامن؟	How many non-permanent members of the Security Council?
D 0153 AR AR	ما هو السي++؟	What is C ++?
D 0161 AR AR	ما هي صنداي تايمز؟	What is the Sunday Times?
D 0165 AR AR	ما هو حزب الاتحاد الاشتراكي للقوات الشعبية؟	What is the Socialist Union of Popular Forces?

[10] http://users.dsic.upv.es/~ybenajiba/downloads.html
[11] http://users.dsic.upv.es/~ybenajiba/

D 0167 AR AR	ما هي سيليكون فالي؟	What is Silicon Valley?
D 0173 AR AR	من هي زوجة محمد رضا البهلوي؟	Who is the wife of Mohammad Reza Pahlavi?
D 0174 AR AR	من أحرز على جائزة نوبل للسلام عام 2002؟	Who won the Nobel Peace Prize in 2002?

Table 5: Questions of Benajiba Dataset.

3.2.1. Criticizing of Benajiba Dataset:

- Answers to old questions.
- Most of the questions are so old that we can‘t get answers.
- Fault writing of named entities.
- Questions translated from foreign cultures.
- Decoding of character of Dataset.

And for the type of WH Question used in Benajiba Dataset , we presnt in the table below:

WH Question	Traduction
كم	How many
في اي	In any
ما	What
على اي	At what
اين	Where
متى	When
من	Who
كيف	How
عن	Which /What
الى من	To who
اي	Which

Table 6: WH Question type of Benajiba Dataset.

3.2.2. Question part:

The questions in our database consist of 4 parts: question words, stop words, central words, prepherical words.in this section we will give some exmple from Benajiba Dataset.

And we will represent this part in the examples below:

How many are the annual joys of Google? ؟كم هي العوايد السنوية لغوغل D 0001 AR AR

؟كم ينفق على التدخين سنويا في الولايات المتحدة الامريكية D 0002 AR AR

How much is spent on smoking annually in the United States?

How long is the rib bath pool? ؟كم يبلغ طول ضلع بركة الحمام D 0004 AR AR

Questions	Words	Traduction	Type
Question 01	كم	How much	Questions words
	هي	Are	Stop words
	العوايد	Joys	Central words
	السنوية	Annual	Prepherical words
	ل	Of	Stop words
	غوغل	Google	Central words
Question 02	كم	How much	Questions words
	ينفق	Spent	Central words
	على	On	Stop words

	التدخين	Smoking	Central words
	سنويا	Annually	Prepherical words
	في	In	Stop words
	الولايات المتحدة الامريكية	United States	Central words
Question 03	كم	How	Questions words
	يبلغ	Reaches	Prepherical words
	طول	Long	Prepherical words
	ضلع	Rib	Prepherical words
	بركة	Bath	Central words
	الحمام	Pool	Central words

Table 7: Example of Question part from the used Dataset Questions.

3.3. Use of Farasa Dataset:

When we saw the work of "Salvatore Romeo et al", which is titled "Language Processing and Learning Models for Community Answering Question in Arabic" [Romeo et al, 2017].

They focused on the problem of question ranking in community question answering (cQA) forums in Arabic. By using word2vec on Farasa Dataset [43], and the results are strong.

When we wanted to create our model with Farasa Dataset, after the test of the results was very far and they don't have the same meaning, also it gave bad results for example:

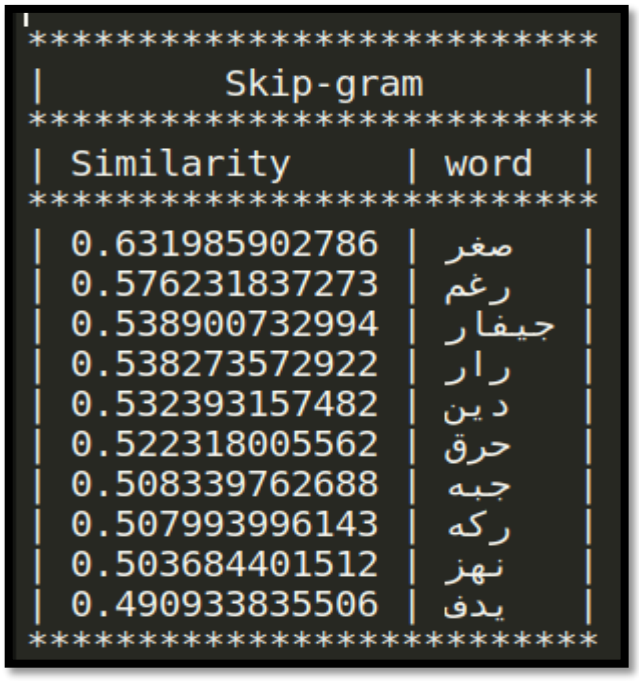


Figure 40: Example for the result of Farasa Dataset.

Then, the number of vocabulary is 7078 words. Then, many words in Benajiba Dataset don't exist in Farasa vocabulary.

3.4. Unique encoding of texts:

All the texts of our Dataset are represented with a UTF-8 which is The Unicode Transformation Format is also known as the Universal Transformation Format.

We used UTF-8 encoding because this encoding supports the Arabic language and we can treat the Arabic language from it.

3.5. Unicodedata:

For this step, we used Unicodata cause of Benajiba Dataset that contains non-normalized characters but with different code ASCII, so to normalize these characters we use Unicodedata.

And for example:

from unicodedata import normalize
print '%r' % normalize ('NFKD', u'\u2460')

3.6. Question reformulation process:

3.6.1. Normalization:

As we said before, the Normalization is the process by which you can perform certain transformations of text to make it reconcilable in a way which it may not have been before.

In part normalization, we worked with the same standardization Affiliate for AraVec models [44] to obtain the same results of its normalization, and this algorithm of normalization:

```
def clean_str (text): #Option1--> normalizing
search = [u"أ",u"إ",u"آ",u"ة",u"_",u"-",u"/",u".",u"،",u" و ",u" يا ",u"",u"-",u"""",u"ى",u"\\",u'\n', u'\t',u'&quot;',u'?',u'؟',u'!']
replace = [u"ا",u"ا",u"ا",u"ه",u" ",u" ",u"",u"",u"",u" و",u" يا",u' ',u"",u"",u"ي",u"",u' ', u' ',u' ',u' ',u' ! ']
 text=araby.normalize_ligature(text)
 text=unicodedata.normalize('NFKD',text)
 text=araby.strip_tashkeel(text) #remove tashkeel
 p_longation = re.compile(r'(.)\1+')#remove longation
 subst = r"\1\1"
 text = re.sub(p_longation, subst, text)
 text = text.replace(u'وو', u'و')
 text = text.replace(u'يي', u'ي')
 text = text.replace(u'اا', u'ا')
 for i in range(0, len(search)):
 text = text.replace(unicodedata.normalize('NFKD',search[i]),
 unicodedata.normalize('NFKD',replace[i])) #trim
 text = text.replace(u'ئ', u'ئ')
```

```
    text = text.strip()
    return text
```

And an example of normalization is illustrated as follow:

بدا→ بدأ(Start)

الافريقيه→ الإفريقية(African)

3.6.2. Skip words between quotation marks:

For all the words in the quotation marks, as we said before, because the words between quotation marks when put them in a search engine, so the search engine looks on an exact match for these words. So, can't extract similar words for them.

Example of these words: "Olive Branch", "غصن الزيتون" ,"African Queen", "الملكة الافريقية","Islamic Mosque Foundation""مؤسسة المسجد الاسلامي"

3.6.3. Remove "Stop words":

Our use of "stop words" is to remove these words because commonly used words that are excluded from searches to help index and parse web pages faster. Some examples of stop words are: "a", "and", "but", "how", "or", and "what." While the majority of all Internet search engines utilize stop words, they do not prevent a user from using them, but they are ignored.

For example, if you were to search for: "What is a motherboard?" on Computer Hope, the search engine would only look for the term "motherboard." [45]

We worked with Arabic stop word list[12], which contain:

Names, Letters, Acts, Conscience, Circumstance, for each type we explain 3 or 2 content.

For Names, we have:

- Name of the signal: There هناك, So هكذا , This هذا
- Connected name: التي Which,الذي Which

For Letters, we have:

- Prepositions:إلى to,على On
- Alphanumeric characters:و And
- Letters: Aأ,Bب

For Conscience, we have:

- A separate Conscience:ب B,ل L
- Connected Conscience:نحن We, they هم

For Circumstance, we have:

- Place Circumstance: Betweenبين , underتحت
- Time Circumstance: Before قبل, whenعند

For Acts, we have:

- Affirmations: Noكلا ,All كل
- Explanatory actions: Evenحتى [46]

[12] https://github.com/linuxscout/arabicstopwords

3.6.4. Skip Named entities:

We have extracted all the named entities exist in benajiba dataset, and also we decided to fix these words it means we don't extract similar words for these them, because they are central words in the questions (see Question part in chapter4), when apply the extraction of similar words on these words it gives other words different in meaning than the original, and for example of Named entities:هوكايدو Hokkaidoفؤاد المهندس, Fouad Engineer,غصن الزيتون olive branch.

3.6.5. Put Named Entities between quotation marks:

Every composed Named Entities we put into quotation mark, for example: غصن الزيتونolive branch.

3.6.6. Transformation number to string (Int to Str):

We convert all the numbers exist in dataset to characters (integer a string), to obtain the numbers in two states (integer and string). This work is based on the use of the "Pyarabic library".

And the structure and as follows, this library search on numbers if it finds a number so it transforms to a string. [13][47]

Example:

23 → (23 | ثلاث و عشرون)

3.6.7. Transformation string to number (Str to Int):

For this part we do the opposite to the previous step, it means we transform the number in character format to integer format; we used the same library "Pyarabic library", and for the structure just library search on numbers in format of the characters, after it transforms to integer.

Example:

1985→ (1985 | ألف و تسعمئة و خمس و ثمانون)

3.6.8. Extract Similar Words to the rest words.

To extract similar words for all the words that remain (Prepherical words), we use the library genism to download the worde2vec models, so it contains all the functions related to the word2vec so this library is the responsible on the extraction of similar words, for each word, it extract 10 similar words.

We use two word2vec models created by: Abu Bakr Soliman, Kareem Eissa, and Samhaa R. El-Beltagy1 presented in the article titled by: "AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP".

[13] https://github.com/linuxscout/pyarabic

3.7. Models of Word2Vec:

For this part, we used one of the two word2vec models, as we said before from this models created we choose to work with the models of Wikipedia exist in first model, Because it contains articles easy to handle, and looking at our dataset questions, then Wikipedia model are the appropriate ones with these questions.

Each model contains 300 dimensions, for CBOW models of Wikipedia contains 1,800,000 Docs and 140,319 vocabulary and Skip Gram models of Wikipedia also contains 1,800,000 Docs and 140,319 vocabularies .

Both models are created from AraVec Project, the creator of this project are: Abu Bakr Soliman, Kareem Eisa and Samha R. El-Bellay, in the Proceedings of the 3rd International Conference on Computational Arabic Linguistics (ACLing 2017), Dubai, WATER, 2017. [44]

AraVec is a pre-trained distributed word representation (word embedding) open source project which aims to provide the Arabic NLP research community with free to use and powerful word embedding models, this project has 2 versions:

The first version of AraVec[14] provides six different word embedding models built on top of three different Arabic content domains; Tweets, World Wide Web pages and Wikipedia Arabic articles. The total number of tokens used to build the models amounts to more than 3,300,000,000. [Abu Bakr Mohammad et al, 2017]

The second version of AraVec[15] provides twelve different word embedding models built on top of three different Arabic content domains; Tweets, World Wide Web pages and Wikipedia Arabic articles. The difference between this version and the first is that the hyper-parameter for minimum count was reduced to 50 instead of 500 for Tweets dataset, 200 for World Wide Web pages dataset and 5 for Wikipedia articles dataset. This resulted in models that have more coverage in terms of vocabulary. The other change is that we produced a set of six embedding models that have a dimension of 100.

The actors wrote an article“AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP” [Abu Bakr Mohammad et al, 2017] explains how to build the 3 models: 2 tweets, 2 World Wide Web, 2 Wikipedia. To know the method of creation you can see the article.

3.8. Similarity:

Each similar word has a value of similarity between it and the origin word.

To calculate the similarity, we use the genism library which is responsible for extracting similar words, which extract for each word 10 similar words with its similarity values.

Genism uses the cosine function to calculate the similarity by default.

[14] https://github.com/bakrianoo/aravec/tree/master/AraVec%201.0
[15] https://github.com/bakrianoo/aravec/blob/master/README.md

3.6.9. Stemming for similar words:

As we said before, Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming is important in natural language understanding (NLU) and natural language processing (NLP). [24]

To do our stemming, we used light stemming by using custom .stemmer from tashafyne.

We use stemming because google search engine compare the path of characters, and if we take as an example:

و الحياة → حياة (And life → life)

So, حياة give many of the results compare to و الحياة

In this step we get reformulated question.

3.9. Get answers of questions:

To extract the answers to the questions (initial and reformulated), we used google API, because google API and the first search engine (number one).

We extracted for each question the first 10 link get through google API which contains answers, the majority of the number of answers graduated incomplete from Google, all the results we have stored in a CSV file.

3.6.10. Determination the threshold of similarity:

To find just the words that have the same meaning as the word origin, we use an interval contains two values(similarity threshold for central words and similarity threshold for prepherical words), so by this values we leave all similar words that have a similarity values more than or equal to the value of the interval we gave.

Then, after each choice of the value of interval we make the test of result, which means we can control this value until we obtain the closest similar words. And to understand this step we have written an algorithm:

Input: (similar words, similarity value), interval.
Output: the closest similar words. The application of this speech is as follows:
most_similar = model.wv.most_similar(araby.normalize_ligature(word))

most_similar= [word for word,similarity in most_similar if similarity>val_sim]
For example:
The results of the closest similar words from CBOW:
D AR AR 0001 (العوائد | عائدات | ارباح | نفقات | سيول | ايرادات | مخصصات | تكاليف | قروض | مداخيل | مكاسب) السنويه لغوغل

The results of the closest similar words from Skip-Gram:
D AR AR 0001 (العوائد | ارباح | عوائد) السنويه لغوغل

3.6.11. Use of “+”sign or “ ” (Space) sign:

After the extraction of similar words, we put the sign "|" between the word origin and similar words and also between similar words. And the “” sign or the “+” sign to indicate the begining of new word. To understand the function of them (see chapterIV).

For example the results and like this, since we have two model of Word2Vec, we present the results of the two models:

Results by CBOW Models by the use of the “+” sign:

D AR AR 0001كم+هي+(العوائد | العائدات | الارباح | النفقات | السيوله | الايرادات | المخصصات | التكاليف | القروض | المداخيل | المكاسب)+(السنويه | السنوي | بالذكري | المئويه | سنويه | سنويا | المقرره | الشهريه | لتاسيسها | سنوي | التحضيريه)+لغوغل

Results by CBOW Models by the use of the “” sign:

D AR AR 0001كم هي (العوائد | عائدات | ارباح | نفقات | سيول | ايرادات | مخصصات | تكاليف | قروض | مداخيل | مكاسب) السنويه لغوغل

Results by Skip-Gram Models by the “+” sign:

D AR AR 0001كم+هي+ (العوائد | والارباح | والعوائد | الارباح | والنفقات | العائدات | والمدفوعات | وارباح | والديون | الفوائض | والاقتراض) + (السنويه | السنوي | سنويه | سنويا | جائزتها | سنوي | والسنويه | لتاسيسها | المئويه | والذكري | المتساقطات) +لغوغل

Results by Skip-Gram Models by the “” sign:

D AR AR 0001 (العوائد | ارباح | عوائد) السنويه لغوغل

3.10. Test results (relevant/irrelevant):

Now each question has 10 links answers retrieved from Google API.

To test the results if it has a relationship with the question (it contains the answer on the question or not).

To do the test we have based on the description of the links, the result is relevant if it answers the question as a percentage of 50%

3.11. Test results (improvement / no improvement):

To measure our results in relation to initial query, we calculate:

- For each query we retrieve the number of relevance.
- In the second phase, we calculate the margin (difference) between the number of relevance for each initial question and that corresponds in query reformulated.
- we group the margins in different lists, a list for the questions that is have a result equal to 0, other equal to 1, then equal to 2(bigram), etc.
- After the extraction of the previous lists we calculate the average of each list. By the following formula (1):

$$M = \text{average}(\text{margin}(\text{initial test}, \text{test(i)}) \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

4. Work presentation:

For the implementation of our approach which is mentioned in (chapter IV). We followed the steps one by one using python software.

4.1. Flow:

Our work is based on the step question reformulation process much more.

We made several tests each time we use some phases of the question reformulation process with each model of Word2Vec.

Tests	CBOW	Skip- Gram	Normalization	Quotation marks	Remove Stop Words	Fix Named Entities	Int2 str	Str2int	Similar Words	Value_Centrale Similarity	Value_Prepherical Similarity	Stemming	Named Entities between ‘ ’
Test01	0	0	0	0	0	0	0	0	0	0	0	0	0
Test02	1	0	1	1	1	1	0	0	1	0,4	0,5	1	0
	0	1	1	1	1	1	0	0	1	0,4	0,5	1	0
Test03	1	0	1	1	1	1	0	0	1	0,6	0,7	1	0
	0	1	1	1	1	1	0	0	1	0,6	0,7	1	0
Test04	1	0	1	1	1	1	0	0	1	0,7	0,7	1	0
	0	1	1	1	1	1	0	0	1	0,7	0,7	1	0
Test05	0	1	1	1	1	1	0	0	1	0,7	0,7	1	1
	1	0	1	1	1	1	0	0	1	0,7	0,7	1	1
Test06	1	0	1	1	1	1	1	0	1	0,7	0,7	1	0
	0	1	1	1	1	1	1	0	1	0,7	0,7	1	0
Test07	1	0	1	1	1	1	0	1	1	0,7	0,7	1	0
	0	1	1	1	1	1	0	1	1	0,7	0,7	1	0

Table 8: Table of Tests.

4.1.1. Test 01:

The first test based on the initial queries, in this test we did not use any reformulation process steps, the purpose of this test in order to compare the results of initial queries and reformulated queries.

4.1.1.1. Results:

The results are presented in the following graph:

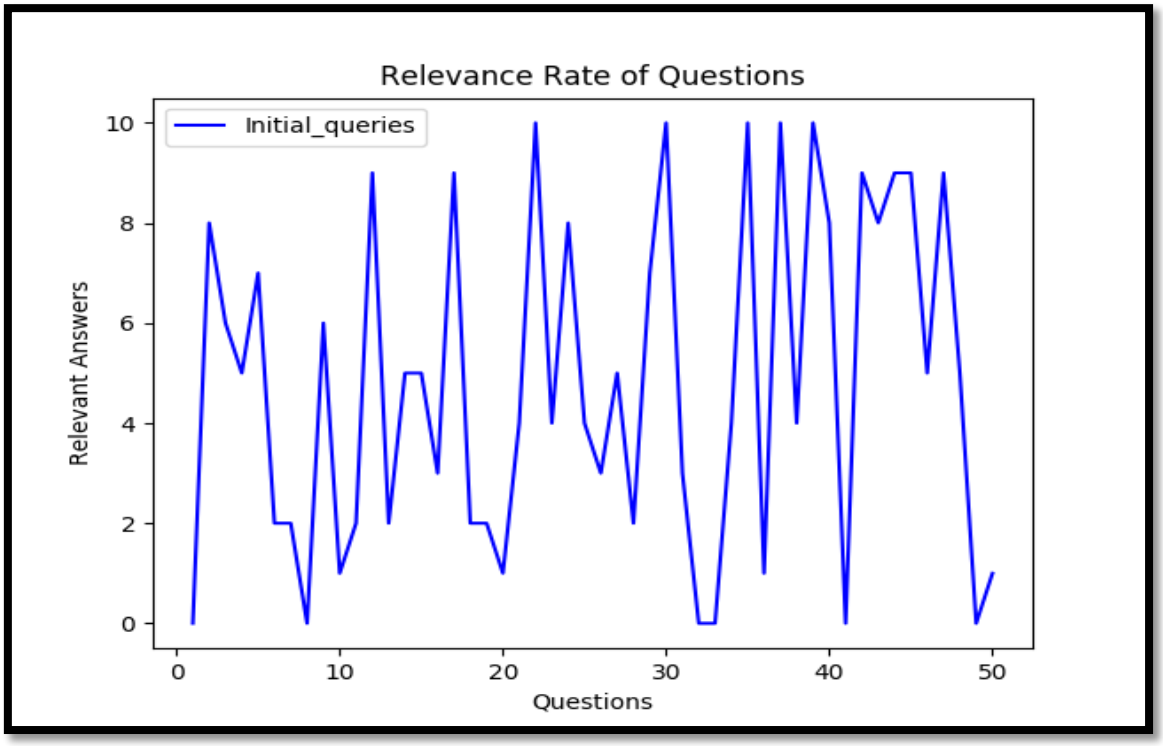


Figure 41: Result of initial queries (without reformulation).

The graph contains the result of relevant answers for the initial queries, this graph is represented by "Line chart" and on the X-axis we find the number of questions from 1 to 50 questions, and on the Y-axis we find the number of relevant answers.

The graph shows us that:

- There are 6 queries without relevant answers 0/10.
- There are 5 queries with 10/10 relevant answers.
- Other queries ranging between 1/10 to 9/10 relevant answers.

Furthermore, the relevant answer percentage for all questions equals 47%.

First, for the queries without any relevant answers which are:

D 0001 AR AR كم هي العوائد السنوية لغوغل؟ How much is Google's annual revenue?
D 0008 AR AR كم عدد موتى اعصار 18 سبتمبر 1906 بهونج كونج؟
How many died of typhoon September 18, 1906 in Hong Kong?
D 0032 AR AR في اي ولاية توجد والا والا؟ In which state are there Walla Walla?
D 0033 AR AR في اي ولاية توجد بلدة ماتيس؟ In what state is Matisse located?
D 0041 AR AR متى اسست الشركة الوطنية الفرنسية للسكك الحديدية؟
When was the French National Railway Company established?
D 0049 AR AR في اي معركة قتل تشارلز الاصلع؟
In what battle did Charles the bald man die?

Second, for the queries with 5 relevant answers are:

D 0004 AR AR كم يبلغ طول ضلع بركة الحمام؟ How long is the rib bath pool?
D 0014 AR AR كم عدد الاعضاء الغير الدائمة في مجلس الامن؟
How many non-permanent members of the Security Council?
D 0015 AR AR كم عدد اعضاء مجلس الامن؟
How many members of the Security Council?

D 0027 AR AR في اي دولة تولى الشعلان منصب وزير الدفاع؟
In what country did Shaalan assume the position of Minister of Defense?
D 0046 AR AR متى انضمت سلطنة عمان الى جامعة الدول العربية؟
When did Oman join the League of Arab States?

Third, for the queries that have relevant answers between 9 and 10 which are:

D 0012 AR AR كم تبلغ قوة انفجار قنبلة نووية؟ How powerful is a nuclear bomb?
D 0017 AR AR ما هي المدة التي يستغرقها كوكب عطارد في دورة واحدة حول الشمس؟
How long does Mercury take in one cycle around the Sun?
D 0022 AR AR كم من كيلوبايت في ميجابايت واحد؟ How many kilobytes per megabyte?
D 0030 AR AR اين ولد ابن بطوطة؟ Where was Ibn Battuta born?
D 0035 AR AR في اي دولة تولى سيرجي لافروف منصب وزير الخارجية؟
In which country did Sergei Lavrov assume the post of Foreign Minister?
D 0037 AR AR متى استشهد عبد القادر الحسيني؟ When did Abdul Qadir al-Husseini die?
D 0039 AR AR متى ولد المطرب كاظم الساهر؟ When was the singer Kazim El Saher born?
D 0044 AR AR متى بدأ لينين الثورة البلشيفية؟
When did Lenin begin the Bolshevik revolution?
D 0045 AR AR متى اكتشف حجر رشيد ابان؟ When was Rosetta stone discovered?
D 0042 AR AR متى اغتيل عبد العزيز الرنتيسي؟
When was Abdel Aziz Rantisi assassinated?
D 0047 AR AR ما هي الرياضة التي يمارسها رونالدو؟ What is Ronaldo's sport?

4.1.1.2. Discussion:

1) After read and analysis the results of queries that have 0/10 and 5/10 relevant answers, we could cite some disadvantages:
   - Different terminologies, some words in the collection are not used a lot in the articles.
   - Old news, use questions found in Wikipedia.
   - Western culture incomplete in Arabic, Use foreign words.
   - Writing errors, for example "والا والا" write adjoining and "وجالينوس" is writing without"و".
   - Stop words still exist.
   - Ambiguity, According to the results gave by Google engine, there are some words have other meaning from what's queries about.
     Ex1: following the results of query "D 0033 AR AR"
     The query goal is to find the place of "Matisse" (بلدة ماتيس) but Google engine gave us result of general "MATISSE (الجنرال ماتيس).
     Ex2: the word "الحمام" can take the meaning of animal and the meaning of bathroom.
   - Split of Named Entities.
     Ex1: "والا والا" its name of city and it's composed of two words, and following to the results of query "D 0032 AR AR" Google engine divide "والا والا" to two words "والا" and " والا" , after that it will gave us results that's have string "والا والا" or "والا".

2) After read and analysis the results of queries that have 9/10 and 10/10 relevant answers, we could cite some advantages:
   - The queries speak on much known topics (scientist, people known, Palestine).

4.1.1.3.Suggestions:

In the end of test, as suggestions there's:

- Add similar words to Prepherical words, with determine threshold of similarity for similar words.
- Remove stop words and WH tools.
- Add Quotations marks between named entities that have more than one word to not be divided.

4.1.2. Test 02:

The second Test based on the initial queries and the reformulated queries, in this one we use the two word2vec models each time, and each models alone, in addition to the normalization, remove stop words, fix Named Entities, extract similar words, stemming the similar words, determine the threshold of similarity, and take just similarities between this last and 1.0: (value similarity for central words= 0.5 and value similarity for Prepherical words = 0.4). We have set the interval to get similar words that have the same meaning with the origin word.

4.1.2.1. Results:

The results are presented in the following graph:

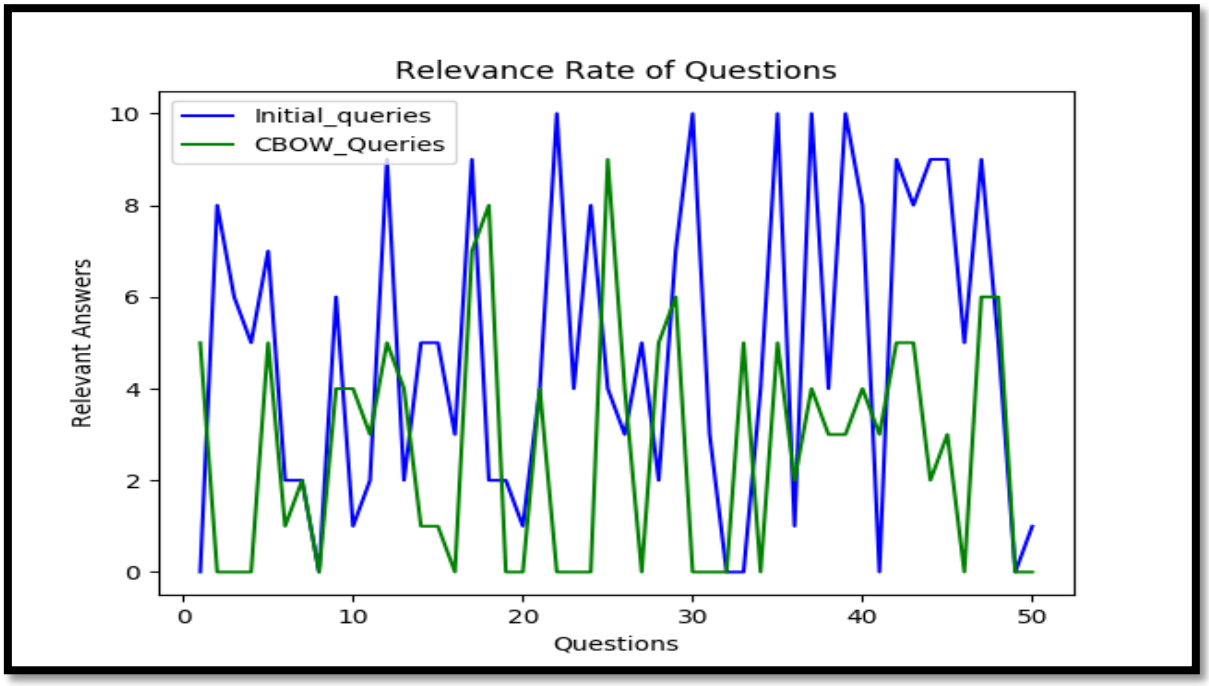


Figure 42: Result of relevant answers by CBOW for test 02.

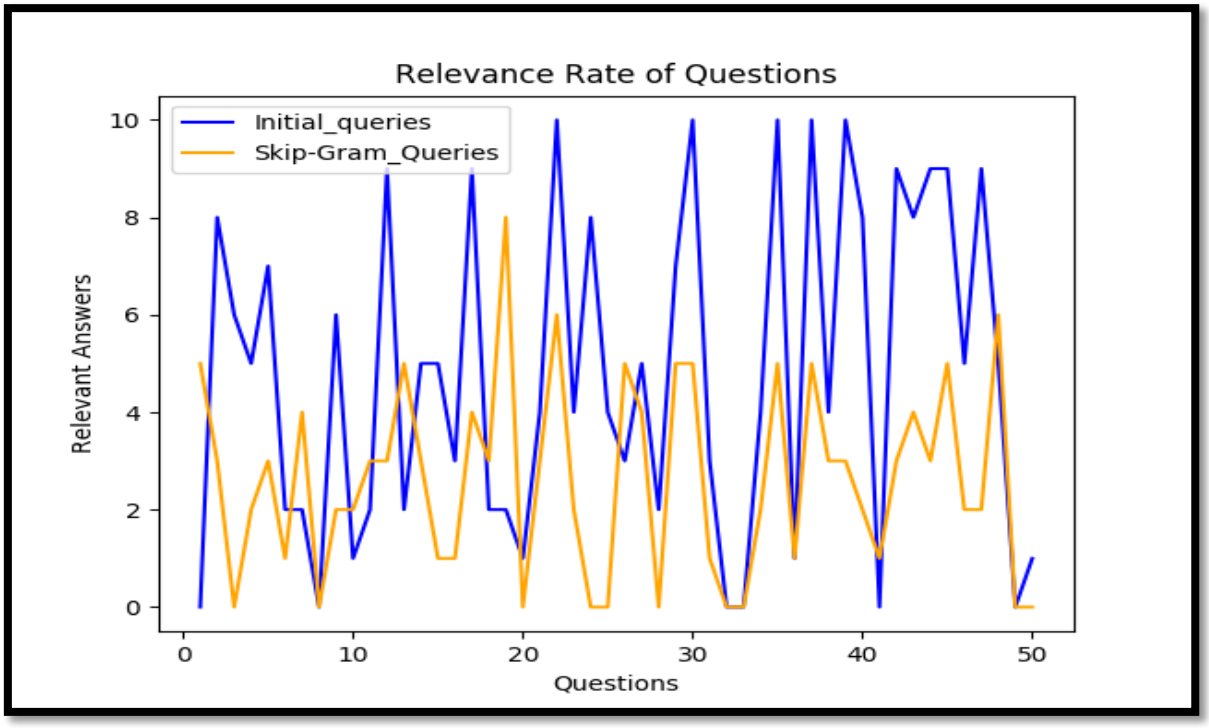


Figure 43: Result of relevant answers by Skip-Gram for test 02.

The two graphs contain the result of relevant answers, for CBOW model in green color (figure42), and the skip-gram model in orange (figure43) and the initial queries in blue with both. This graphs are represented by "Line chart", on the X-axis we find the number of questions from 1 to 50 questions, and On the Y-axis we find a number of relevant answers.

4.1.2.1.1. Results of CBOW model:

From the first graph (Figure42) we see that the results for the queries without any relevant answers are very large which equal to 18 queries in comparison with the queries that have relevant answers 8 or 9 which equal to 2 queries.

Also we see that the results for the queries that have 5 relevant answers its equal to 9 queries, rather the percentage of queries with CBOW model is (26%) very little than the percentage of initial queries relevance results (47%).

4.1.2.2. Discussion of CBOW model:

The percentage of queries with CBOW model (26%) very little from the percentage of initial queries (47%).The difference of percentage return for several problems:

- The interval takes small values for that there are similar words that are very far in meaning from the queries.
- Several questions haven't similar words.

We used this table to clearly specify the results:

The table shown number of reformulated queries for each value of reformulated words (uni gram, bigram, trigram, etc.), and the percentage of this queries reformulated by CBOW with a justification.

- **NRW=Number of Reformulated Word.**

- **NQ=Number of Questions.**
- **RA=Relevant Answers.**
- **I/D=Improvement/diminution.**

NRW	NQ	RA %	+I/-D	Justification
0	1	+10.0	+I	1) For the value 0, there is only one question underwent an improvement of 1 relevant answer more than the initial queries as a consequence of removing the stop words, this improvement to a percentage of 10% for the two word2vec models, the query is: "D AR AR 0048 جنسيه+عباس+الموسوي". 2)There are questions that have the most similar words have relation and meaning with origin word like: Ex1 : D AR AR 0001 العوائد | ارباح | عوائد | نفقات | عائدات | مدفوعات | ديون | فوائض | اقتراض) +(السنويه | سنو | سنوي | سنويا | جائزت | تاسيس | مئوي | ذكر | متساقطات +(لغوغل) Ex2 : D AR AR 0019 عدد | عددا | عديد | مجموع) + (انواع | اصناف | انوع | نوع | اشكال | انواعا +(السلحفيات)
1	9	-7.78	-D	1) Many similar words in these queries haven't the same meaning such a relation with the origin word. Ex1: D AR AR 0009 (كانت | ظلت | قيت | اصبحت | صارت | اتت | اعتبرت | انت | ستكون | اضحت | دت)+(قوه | قو | قوت | قدرات)+(زلزال | اعصار | تسونام | حريق | ثوران | انفجار)+هوكايدو 2) There is a reformulation also at the level of central word with similar words, and also with bad percentage of similar words and that make the meaning lost. 3) Devision of Named Entities 4) Lost the main meaning in two queries when we removed stop words : Ex1: D AR AR 0011 (كتاب | شرح ...)+(الفه |...) + وجالينوس+(الطب |...)+(والصيدله |...) The initial query asks about the number of books in medicine pharmacy but here ask about book in last domains. 5) Ambiguities as we explain in test 01.
2	13	-4.62	-D	
3	9	-2.22	-D	
4	11	0	Null	
5	5	-10.0	-D	
6	2	-10.0	-D	

Table 9: Percentage of relevant answers for CBOW model and their justification (Test 02).

4.1.2.1.2. Results of Skip-Gram model:

From the second graph(Figure43) we see that results for the queries without any relevant answers are very large which equal to 10 queries and the queries that have relevant answers 8 or 9 its equals to 2 queries.

Also we see that the results for the queries that have 1 or 2 relevant answers equal to 15 queries, rather the percentage of queries with Skip-Gram model is (26%) very little than the percentage of initial queries relevance results (47%).

4.1.2.3. Discussion of Skip-Gram:

The percentage of queries results reformulates by Skip-Gram model (26%) very little from the percentage of initial queries results (47%), and this return for several problems, this problems are the same problems of CBOW model.

We used this table to clearly specify the results:

The table shown number of reformulated queries for each value of reformulated words (uni gram, bigram, trigram, etc.), and the percentage of this queries reformulated by Skip-Gram with a justification.

NRW	NQ	RA%	+I/-D	Justification
0	1	+10.0	+I	1) For the value 0, there is only one question underwent an improvement of 1 relevant answer more than the initial queries as a consequence of removing the stop words, this improvement to a percentage of 10% for the two word2vec models, the query is: "D AR AR 0048 جنسيه+عباس+الموسوي". 2)There are questions that have the most similar words have relation and meaning with origin word like: Ex1 : D AR AR 0001 (العوائد | ارباح | عوائد | نفقات | عائدات | مدفوعات | ديون | فوائض | اقتراض)+(السنويه | سنو | سنوي | سنويا | جائزت | تاسيس | مئوي | ذكر | متساقطات)+لغوغل Ex2 : D AR AR 0019 (عدد | عددا | عديد | مجموع)+(انواع | اصناف | انوع | نوع | اشكال | انواعا)+السلحفيات
1	9	-7.78	-D	1) Many similar words in these queries haven't the same meaning such a relation with the origin word. Ex1: D AR AR 0009 (كانت | انت | ظلت | تكن | قيت | اصبحت | صارت | ستكون | كانت)+(قوه | قو | قوت | دافع)+(زلزال | تسونام | اعصار | زلازل | زلزالا)+هوكايدو 2) There is a reformulation also at the level of central word with similar words, and also with bad percentage of similar words and that make the meaning lost. 3) Lost the main meaning in two queries when we removed stop words : Ex1: D AR AR 0011 (كتاب |...)+(الفه | ...) + وجالينوس+(الطب | صيدل | طب | جراح | بيطر | كيمياء)+(والصيدله |...) The initial query asks about the number of books in medicine pharmacy but here ask about book in last domains. 4) Ambiguities in the sense as we explain in the first test. 5) Google engine Split named entities.
2	13	-3.85	-D	
3	9	-4.44	-D	
4	11	-2.73	-D	
5	5	-10.0	-D	
6	2	-10.0	-D	

Table 10: Percentage of relevant answers for Skip-Gram model and their justification (Test 02).

Notes:

1) Every time we specify big values for the interval we get few similar words, and every time we specify small values for the interval we get more similar words.

2) Most results obtained from Google was limited from 3 to 6 answers no more because of the large number of similar words given, since most questions have 11 similar words for each question word, and because of "+" sign that worked as logic and.

3) In the end of test we can say there isn't any improvement for (CBOW and Skip-Gram models).

4.1.2.4. Suggestions:

- Raise the value of similarity of similar words for give us the similar words have the same meaning with the origin word.
- Add quotation between named entities, to get an exact result about Name Entities.

4.1.3. Test 03:

The third test based on the second test queries, in this test we use the two word2vec models each time, and each models alone, in addition to the normalization, remove stop words, fix between quotation mark, fix Named Entity, extract similar words, stemming, determine the threshold of similarity, and take just similarities between this last and 1.0: (value similarity for central words= 0.7 and value similarity for Prepherical words = 0.6). We have set values largest than values of test 02 for the threshold to get most similar words that have the same meaning with the origin word.

4.1.3.1. Results:

The results are presented in the following graph:

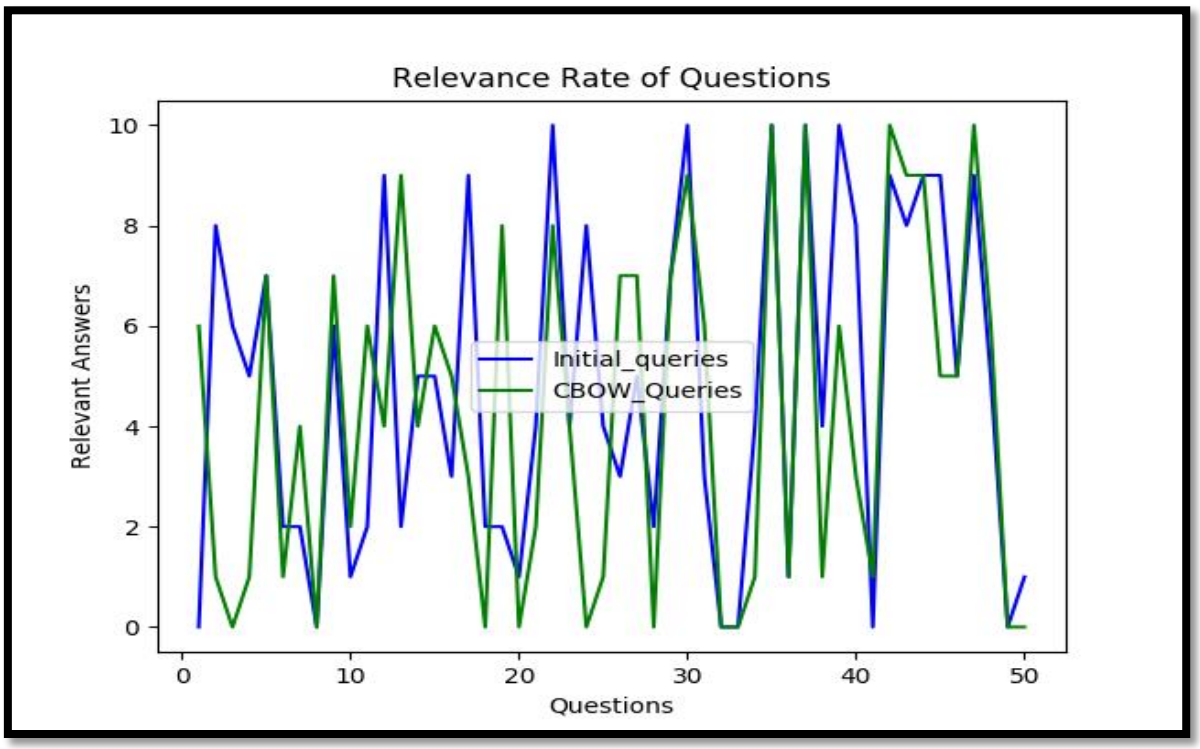


Figure 44: Result of relevant answers by CBOW for test 03.

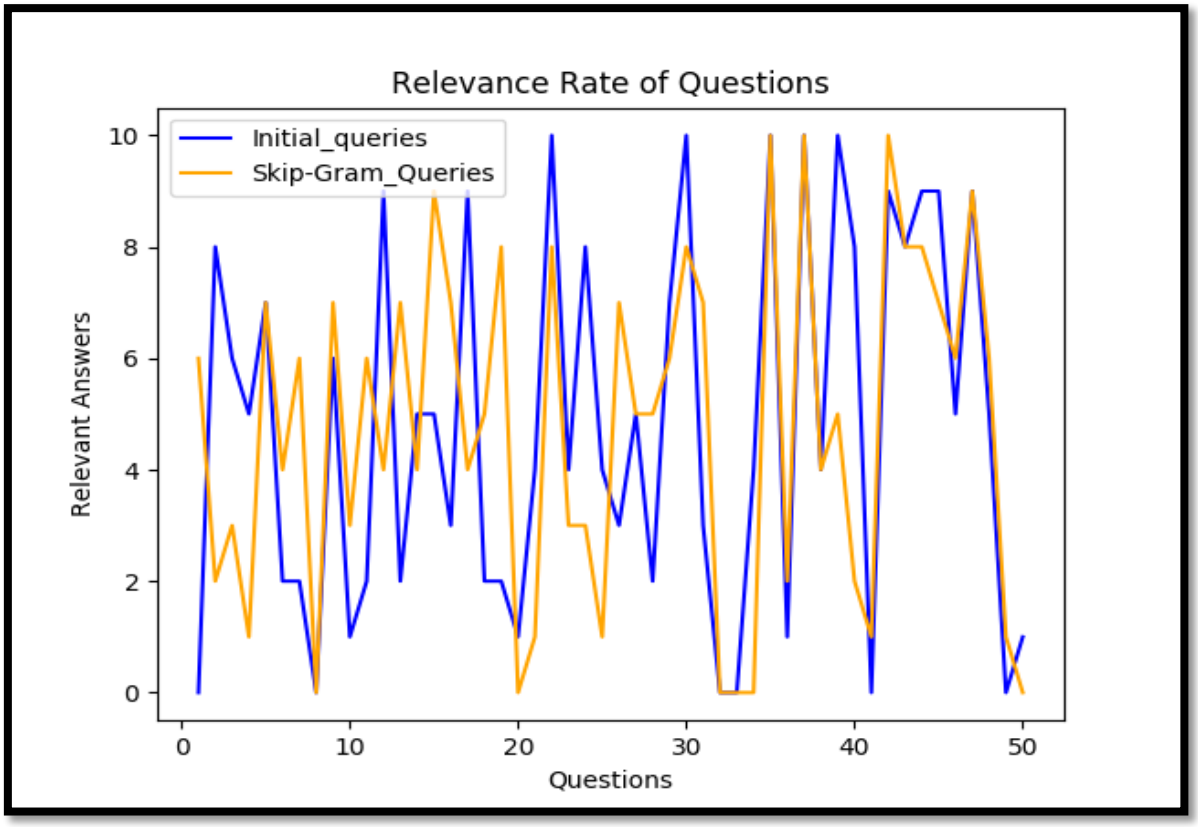


Figure 45: Result of relevant answers by Skip-Gram for test 03.

The two graphs contain the result of relevant answers for CBOW model with green color (Figure44) and for skip-gram model in orange color (Figure45) with blue color for the initial queries. These two graphs are represented by Line chart, on the X-axis we find the number of questions from 1 to 50 questions. And On the Y-axis we find a number of relevant answers.

4.1.3.1.1. Results of CBOW model:

From the first graph (figure44) we see that the results for the queries without any relevant answers are much converged which equal to 10 queries with the queries that have 10 relevant answers equals it's to 4 queries.

Also we see that the results for the queries that have 6 relevant answers which equal to 6 queries, rather the percentage of queries with CBOW model (42%) converged from the percentage of initial queries (47%).

4.1.3.2. Discussion of CBOW model:

The percentage of queries with CBOW model (42%) converged from the percentage of initial queries (47%), and this increase return to the augmentation of interval values [0.6, 0.7], despite that the author steps are even that the test before.

We used this table to clearly specify the results:

The table shown a number of reformulated queries for each value of reformulated words (uni gram, bigram, trigram, etc.), and the percentage of this queries reformulated by CBOW with a justification.

NRW	NQ	RA%	+I/-D	Justification
0	3	+3.33	+I	1) The number of similar words decreases. 2) Most similar words (50%) have the same meaning as the origin words. 3) The raise of the values of similarity in order to reduce the number of similar words extracted for each word.
1	13	+3.01	+I	
2	17	-3.53	-D	1) We have not yet reached the point where most of similar words have relation and the same meaning as the origin word. 2) Lost the main meaning when we removed stop words. 3) Ambiguities in the sense as we explain in the first test. 4) Google engine Split named entities.
3	10	-1.0	-D	
4	5	-4.0	-D	
5	1	-10.0	-D	
6	1	-10.0	-D	

Table 11: Percentage of relevant answers for CBOW model and their justification (Test 03).

4.1.3.1.2. Results of Skip-Gram model:

From the second graph (Figure45) we see that the results for the queries without any relevant answers are converged which is equal to 6 queries in comparison with the queries that have relevant answers 10 its equals to 3 queries.

Also we see that the results for the queries that have 6 relevant answers equal to 6 queries, rather the percentage of queries with Skip-Gram model (47%) equal the percentage of initial queries (47%).

4.1.3.3. Discussion of Skip-Gram model:

The percentage of queries with CBOW model (47%) equal to the percentage of initial queries (47%), and this return to the augmentation of interval values [0.6, 0.7], despite that the author steps are even that the test before.

We used this table to clearly specify the results:

The table shown number of reformulated queries for each value of reformulated words (uni gram, bigram, trigram, etc.), and the percentage of this queries reformulated by Skip-Gram with a justification.

NRW	NQ	RA%	+I/-D	Justification
0	3	+3.33	+I	1) The number of similar words decreases. 2) Most similar words (50%) have the same meaning as the origin words. 3) The raise of the values of similarity in order to reduce the number of similar words extracted for each word.
1	13	+2.31	+I	
5	1	+10.0	+I	
2	17	-0.59	-D	1) We have not yet reached the point where most of similar words have relation and the same meaning as the origin word. 2) Lost the main meaning when we removed stop words.
3	10	-1.0	-D	
4	5	-4.0	-D	

6	1	-10.0	-D	3) Ambiguities in the sense as we explain in the first test. 4) Google engine Split named entities.

Table 12: Percentage of relevant answers for Skip-Gram model and their justification (Test 03).

Notes:

1) Every time we specify big values for the interval we get few similar words, and every time we specify small values for the interval we get more similar words.

2)Most results obtained from Google was limited from 4 to 9 answers no more because of the large number of similar words given, since most questions have 11 similar words for each question word, and because of "+" sign that worked as logic and.

3) In the end of the test we can say there isn't any improvement for (CBOW and Skip-Gram models). But there's improvement according test 01.

4.1.3.4. Suggestions:

- Raise the threshold of similarity, for give us the similar words that have the same meaning and relation with the origin word.
- Add quotation between named entities, to get an exact result about Name Entities.
- The use of Space (" ") sign in place of "+" sign to skip google limitations in answers.

4.1.4. Test 04:

The fourth test based on the third test queries ,in this test we use the two word2vec models each time, and each models alone, in addition to the normalization, remove stop words, fix between quotation mark, fix Named Entity, extract similar words, stemming, and raise the threshold of similarity, to take just similarities between this last and 1.0: (value similarity for central words= 0.7 and value similarity for Prepherical words = 0.7), We have set values largest than values of test 03 for the threshold to get most similar words that have the same meaning with the origin word, and we replace Space(" ") by "+" between origin words.

4.1.4.1. Results:

The results are presented in the following graph:

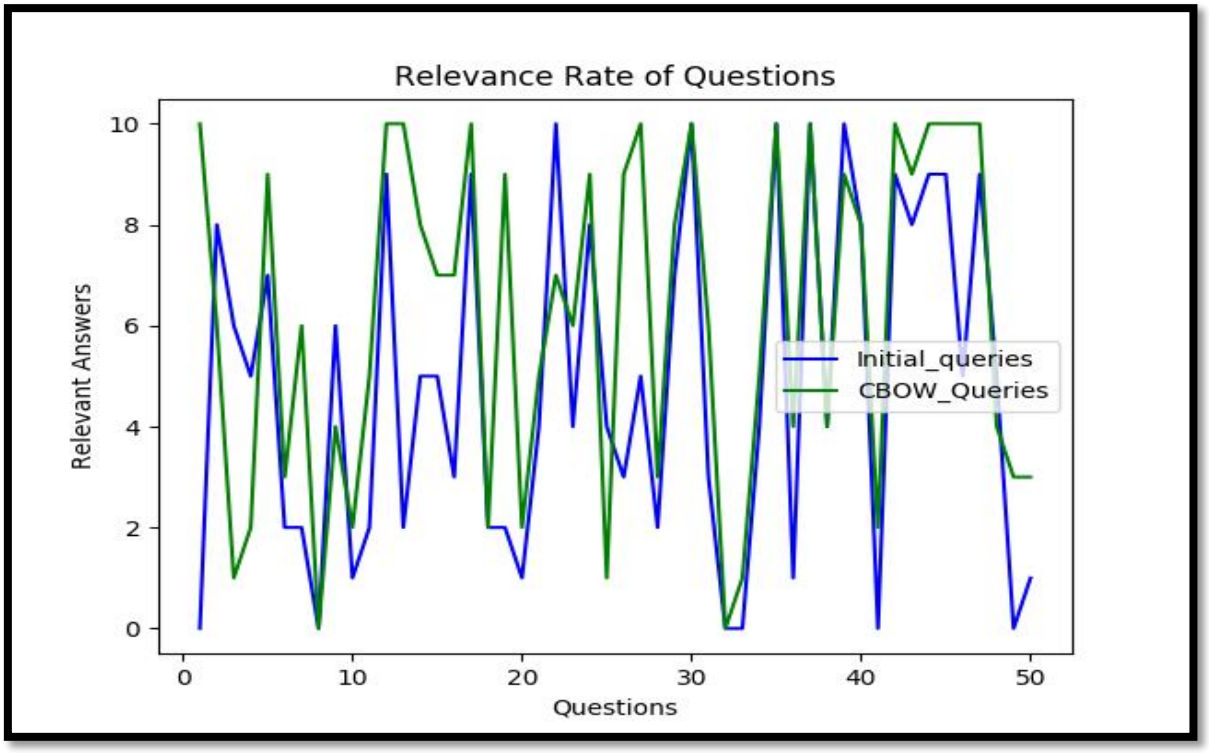


Figure 46: Result of relevant answers by CBOW for test 04.

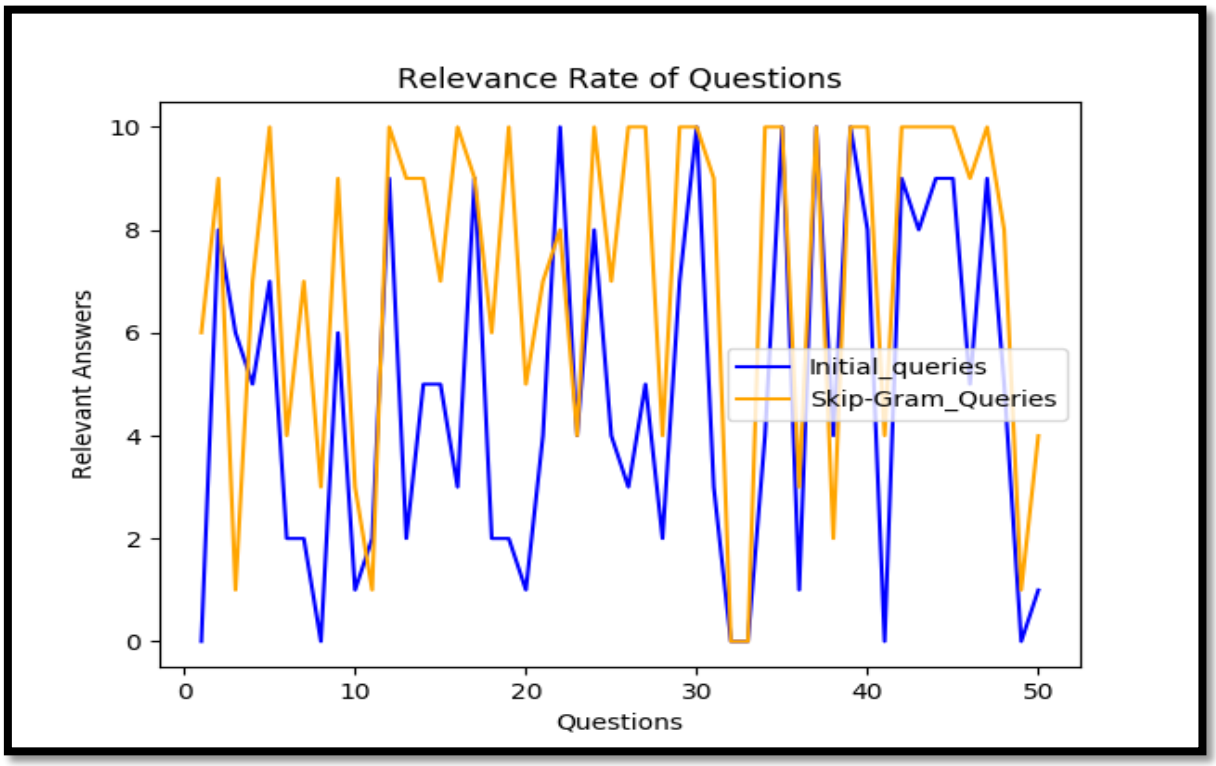


Figure 47: Result of relevant answers by Skip-Gram for test 04.

The two graphs contain the result of relevant answers for CBOW model with green color (figure46) and for skip-gram model with orange color (figure47), with the initial queries. These graphs are represented by Line chart; on the X- axis we find the number of questions from 1 to 50 questions. And on the Y-axis we find a number of relevant answers.

4.1.4.1.1. Results of CBOW model:

From the first graph (figure46) we see that the results for the queries without any relevant answers are very little which equal 10 queries with the queries that have 10 relevant answers equals 11queries.

Also we see that the results for the queries that have 6 relevant answers equal4 queries, rather the percentage of queries with CBOW model (62%) are bigger from the percentage of initial queries (47%).

4.1.4.2. Discussion of CBOW model:

The percentage of queries with CBOW model (62%) are bigger from the percentage of initial queries (47%), and this improvement return also to the augmentation of interval values [0.7, 0.7] and for the replacing the "+" by the Space (" "), despite that the author steps are even that the test before.

We used this table to clearly specify the results:

The table shown number of reformulated queries for each value of reformulated words (uni gram, bigram, trigram, etc.), and the percentage of this queries reformulated by CBOW with a justification.

NRW	NQ	RA	+I/-D	Justification
0	14	+2.86	+I	1) Percentage of improvement is small because we have removed the stop words.
1	20	+5.0	+I	2) Similar words have relation in meaning with the origin words (90%). 3) Number of similar words is small. Similar words have relation with the origin word despite the number of similar words are large.
2	9	+8.89	+I	
3	5	+4.0	+I	
4	2	+10.0	+I	

Table 13: Percentage of relevant answers for CBOW model and their justification (Test 04).

4.1.4.1.2. Results of Skip-Gram model:

From the second graph (figure47) we see that the results for the queries without any relevant answers are converged which is equal to 2 queries in comparison with the queries that have relevant answers 10 its equals to19 queries.

Also we see that the results for the queries that have 6 relevant answers equal to 3 queries, rather the percentage of queries with Skip-Gram model (71%) are big from the percentage of initial queries (47%).

4.1.4.3. Discussion of Skip-Gram:

The percentage of queries with CBOW model (62%) are bigger from the percentage of initial queries (47%), and this improvement return also to the augmentation of interval values

[0.7, 0.7] and to the replacing the "+" by the "", despite that the author steps are even that the test before.

We used this table to clearly specify the results:

The table shown number of reformulated queries for each value of reformulated words (uni gram, bigram, trigram, etc.), and the percentage of this queries reformulated by Skip-Gram with a justification.

NRW	NQ	RA%	+I/-D	Justification
0	14	+6.43	+I	1) There is an improvement because we have removed the stop words.
1	20	+8.0	+I	2) Similar words have relation in meaning with the origin words (90%). 3) Number of similar words is small.
2	9	+8.89	+I	
3	5	-2.0	-D	4)similar words haven't relation with the origin word, for example:.الجائزة : الجوائز, أفضل, جائز..
4	2	+10.0	+I	5) Similar words have relation with the origin word despite the number of similar words are large.

Table 14: Percentage of relevant answers for Skip-Gram model and their justification (Test 04).

Notes:

1) Every time we specify big values for the interval we get few similar words, and every time we specify small values for the interval we get more similar words.

2) In this test the most google results, equal to 10 results for each query.

3) As a result we can say for this test there is an improvement for (CBOW and Skip-Gram models).

4.1.4.4. Suggestions:

- Add quotation between named entities, to get an exact result about Name Entities.
- Add String Format to the Number
- Add Number format for String number

4.1.5. Test 05:

The fifth test based on the fourth test, in this test we use the two word2vec models each time, and each models alone, in addition to the normalization, remove stop words, fix between quotation mark, fix Named Entity, extract similar words, put Named Entities between quotation marks, stemming, replace "+" by " and determine the threshold of similarity as test 04, to take just similarities between this last and 1.0: (value similarity for central words= 0.7 and value similarity for Prepherical words = 0.7).

4.1.5.1. Results:

The results are presented in the following graph:

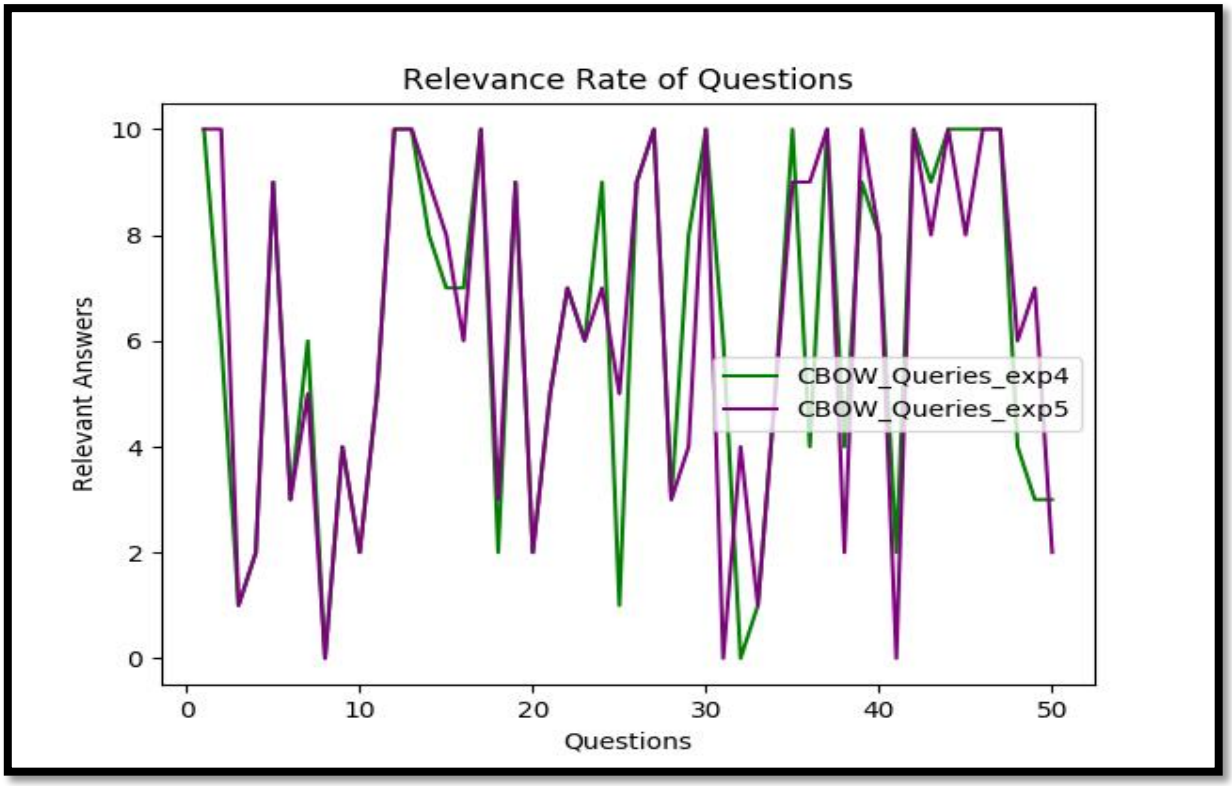


Figure 48: Result of comparison between test04 and test 05 by CBOW.

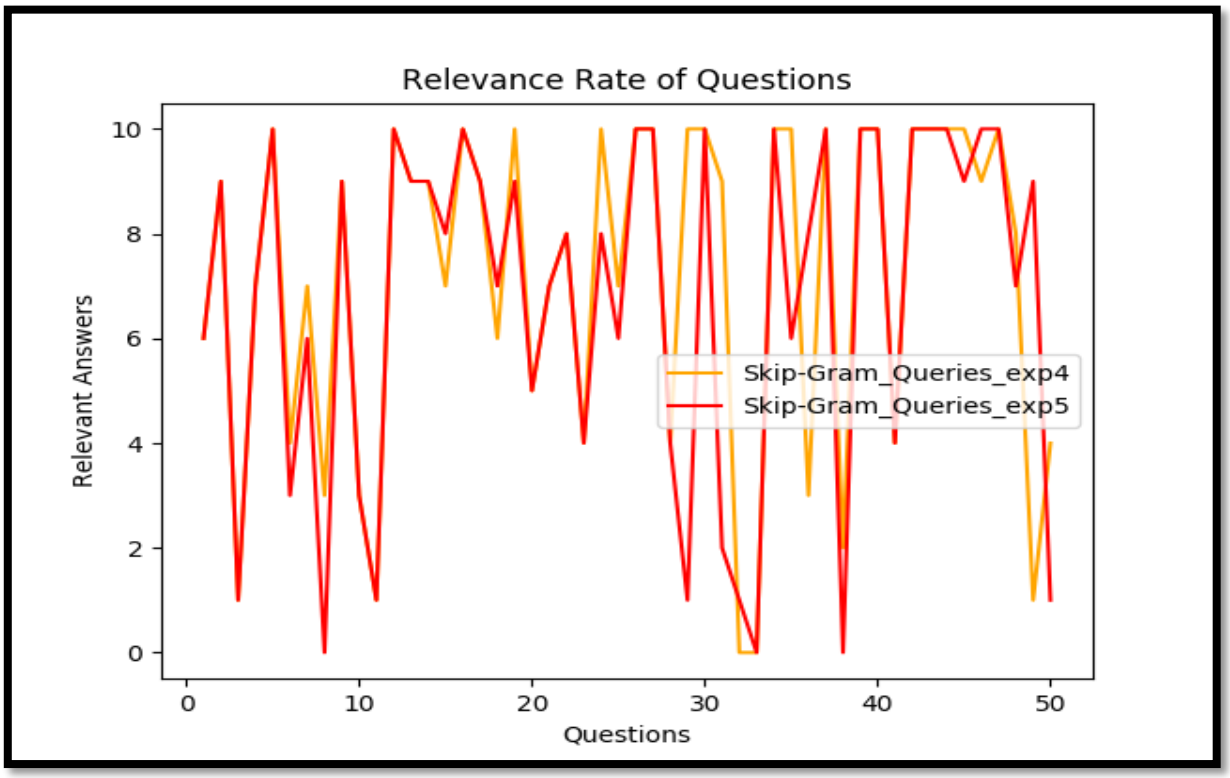


Figure 49: Result of comparison between test 04 and test 05 by Skip-Gram.

The two graphs contain the result of comparison between test 4 and test5 of relevant answers for CBOW model (figure48) and for skip-gram model (figure49), also for the initial queries. This graphs are represented by Line chart, on the X-axis we find the number of questions from 1 to 50 questions. And On the Y-axis we find a number of relevant answers.

For the fifth test we can cite the result below then we discuss it.

4.1.5.1.1. Results of CBOW model:

From the first graph (figure48) we see that the results for the queries without any relevant answers are much converged which equal to 3 queries with the queries that have 10 relevant answers equals to 17 queries.

Also we see that the results for the queries that have 6 relevant answers equal to 3 queries, rather the percentage of queries with CBOW model (67%) are bigger from the percentage of initial queries (47%).

4.1.5.2. Discussion of CBOW model:

The percentage of queries with CBOW model (67%) are bigger from the percentage of initial queries (47%), and this improvement return also to the augmentation of interval values [0.7, 0.7] and replacing the put Named Entity between quotation marks despite that the author steps are even that the test before.

We used this table to clearly specify the results:

The table shown number of reformulated queries for each value of reformulated words (uni gram, bigram, trigram, etc.), and the percentage of this queries reformulated by CBOW with a justification.

NRW	NQ	RA%	+I/-D	Justification
0	18	+5.56	+I	1) Number of similar words is small. 2) Similar words have relation in meaning with the origin word. 3) When we remove the division of Named Entities problem, than we put on searching we get the right answers for the real meaning.
1	18	+7.22	+I	
2	8	+10.0	+I	
3	4	+2.5	+I	
4	2	+5.0	+I	

Table 15: Percentage of relevant answers for CBOW model and their justification (Test 05).

4.1.5.1.2. Results of Skip-Gram model:

From the graph (figure49) we see that the results for the queries without any relevant answers are converged equal to 1 query in comparison with the queries that have relevant answers 10 which is equals to 15 queries.

Also we see that the results for the queries that have 6 relevant answers equal to 3 queries, rather the percentage of queries with Skip-Gram model (69%) are big from the percentage of initial queries (47%).

4.1.5.3. Discussion of Skip-Gram model:

The percentage of queries with CBOW model (69%) are bigger from the percentage of initial queries (47%), and this improvement return also to the augmentation of interval values

(0.7, 0.7) and replacing the put Named Entity between quotation marks despite that the author steps are even that the test before.

We used this table to clearly specify the results:

The table shown number of reformulated queries for each value of reformulated words (uni gram, bigram, trigram, etc.), and the percentage of this queries reformulated by Skip-Gram with a justification.

NR	NQ	RA%	+I/-D	Justification
0	18	+6.67	+I	1) Number of similar words is small. 2) Similar words have relation in meaning with the origin word. 3) When we remove the division of Named Entities problem, than we put on searching we get the right answers for the real meaning.
1	18	+8.34	+I	
2	8	+9.99	+I	
3	4	+4.99	+I	
4	2	+9.99	+I	

Table 16: Percentage of relevant answers for Skip-Gram model and their justification (Test 05).

Notes:

As test 5 is the same as test 4 except that we added "put Named Entity between quotation marks ". When we extract the percentage of answers relevant for the two tests, and for each model (CBOW and Skip-Gram), we make the margin between the relevant percentages of answers of each model. If the result is positive then there is an improvement otherwise there is no.

This compare is done to see if there is an improvement when we add "put Named Entity between quotation marks (“”).

4.1.5.1.3. Results of comparison:

And this table (table17) contains the result of comparison:

Model	NQR By “”	Test04	Test05	+	-	Null	+I/-D
Skip-Gram	25	71%	67%	6	13	6	-D
CBOW	25	61%	61%	11	11	3	+I

Table 17: Result of the comparison.

Query ID	Named Entities	Test 04		Test05		Evaluation	
		CBOW	Skip-Gram	CBOW	Skip-Gram	CBOW	Skip-Gram
D AR AR 0002	الولايات المتحدة الامريكية	6	9	10	9	+I	N
D AR AR 0006	سعود عبد العزيز بن متعب الرشيد	3	4		3	N	-D
D AR AR 0007	كات ستيفنز	7	6	6	5	-D	-D
D AR AR 0008	بهونج كونج	3	0	0	0	N	-D

D AR AR 0012	قنبلة نووية	10	10	10	10	N	N
D AR AR 0014	مجلس الامن	9	8	9	9	+I	N
D AR AR 0015	مجلس الامن	7	7	8	9	+I	+I
D AR AR 0016	السلطان الاشرف خليل	10	7	10	6	-D	-D
D AR AR 0024	فؤاد المهندس	10	9	8	7	-D	-D
D AR AR 0029	تايم وارنر	10	8	1	4	-D	-D
D AR AR 0030	ابن بطوطة	10	10	10	10	N	N
D AR AR 0031	كليف ريتشارد	9	6	2	0	-D	-D
D AR AR 0032	والا والا	0	0	1	4	+I	+I
D AR AR 0035	سيرجي لافروف	10	10	6	9	-D	-D
D AR AR 0036	سلستين الثالث	3	4	8	9	+I	+I
D AR AR 0037	عبد القادر الحسيني	10	10	10	10	N	N
D AR AR 0038	نيويورك تايمز	2	4	0	2	-D	-D
D AR AR 0039	كاظم الساهر	10	9	10	10	+I	N
D AR AR 0041	الشركة الوطنية الفرنسية للسكك الحديدية	4	2	4	0	-D	-D
D AR AR 0042	عبد العزيز الرنتيسي	10	10	10	10	N	N
D AR AR 0045	رشيد ابان	10	10	9	8	-D	-D
D AR AR 0046	سلطنة عمان	9	10	10	10	N	+I
D AR AR 0048	عباس الموسوي	8	4	7	6	+I	-D
D AR AR 0049	تشارلز الاصلع	1	3	9	7	+I	+I
D AR AR 0050	براند نبرغ	4	3	1	2	-D	-D

Table 18: Comparison of Named Entities between Test4 and Test5.

From this table (Table 18) and our vision we have:

- Most of queries which equal 14 queries have (+1) improvement or (-1) diminuation or also no improvement or dimunation.
- We have 4 queries have +2 imrovement because general of Named Entities Unknown For example: "سلستين الثالث", "تشارلز الاصلع", "والا والا".
- 7 queries have diminuation because the most of Named Entities are known.

4.1.5.3. Discussion results of comparison:

From the graph (Figure48), for CBOW model we see the result are convergent, exist some changes like from the question number 20 to the question number 30 we that results of test 4 are higher from the results of test5.

From the graph (Figure49), for Skip-Gram model we see the result are also convergent, exist some changes like from the question number 20 to the question number 30 we that results of test 4 are higher from the results of test5.and from the first question the question number 5 there is only result of test 5 while we don't have any result from the test4.

We used this table to clearly specify the results:

The table shown number of reformulated queries for each value of reformulated words (uni gram, bigram, trigram, etc.), and the percentage of this queries reformulated by Skip-Gram with a justification.

NRW	NQ	RA%	+I/-D	Justification
0	18	-2.14	-D	1) Ambiguities in the sense for example: in the sentence below: متى اصدرت لأول مرة نيويورك تايمز؟ Intended to establish the New York Times and not what the New York Times issued. 2) Number of relevant answers is small.
1	18	-0.95	-D	
2	8	1.25	+I	1) There is an improvement because we have removed the stop words.
3	4	-4.0	-D	1) Ambiguities in the sense for example: in the sentence below: متى اصدرت لأول مرة نيويورك تايمز؟ Intended to establish the New York Times and not what the New York Times issued. 2) Number of relevant answers is small.
4	2	-5.0	-D	

Table 19: Comparison between percentage of relevant answers for test 4 and 5.

Notes:

1) Finally, As a result we can say that there is an improvement of 1% for CBOW model and diminution of 1% for Skip-gram model.

2) We waited a more improvement of results but that's not done, and that return to complicated algorithms of google engine as black bottle.

3) We waited an improvement from Quotation between Named entities, but we didn't get it, and that because of google engine is black buttle.

3) Finally, As a result we can say that there is an improvement for (CBOW and Skip-Gram models).

4.1.5.4. Suggestions:

- Transformation from number to string (Int2str).
- Transformation from string to number (Str2int).
- Change some stop words 2 words as كم من → عدد

4.1.6. Test 06:

In this test we use the two word2vec models each time, and each models alone, in addition to the normalization, remove stop words, fix between quotation mark, fix Named Entity, transformation Str to Int, extract similar words, stemming, replace "+" by " ", and raise the threshold of similarity, to take just similarities between this last and 1.0: (value similarity for central words= 0.7 and value similarity for Prepherical words = 0.7), We have set values largest than values of test 05 for the threshold to get most similar words that have the same meaning with the origin word.

4.1.6.1. Results:

For the CBOW model, there is one query which has 1/10 relevant answers. And for the Skip-Gram model, there is one query which has 1/10 relevant answers.

4.1.6.2. Discussion:

We have set the interval to get similar words that have the same meaning with the origin word; the current results indicate that there is a change in 4 questions from the 50 questions in total and with the use of word2vec Models (one time CBOW and other time Skip-Gram). This change is the transformation of numbers into letters, these questions for CBOW and Skip-Gram models are:

4.1.6.3. Queries transformation for CBOW model:

1) D 0008 AR AR كم عدد موتى اعصار 18 سبتمبر 1906 بهونج كونج؟ --->

D AR AR 0008 عدد موتي (اعصار | زلزال) (ثماني عشرة | 18) سبتمبر (ألف و تسعمئة و ست | 1906) بهونج كونج

2) D 0020 AR AR على اي ارتفاع انفجرت الطائرة الهندية يوم 23 يونيو 1985؟ --->

D AR AR 0020 ارتفاع (انفجرت | تحطمت | اصطدمت | اخترقت | انفجر | قصفت) الطائره الهنديه يوم (ثلاث و عشرون | 23) (يونيو | ديسمبر | مايو | اكتوبر | نوفمبر | يوليو | يناير | سبتمبر | ابريل | مارس | راير) (ألف و تسعمئة و خمس و ثمانون | 1985)

3) D 0028 AR AR ما هي الدولة التي هزمت الدولة العثمانية في 10 يناير 1475؟--->

D AR AR 0028 (الدوله | دول) (هزمت | هزم | خسرت | غزت | انتصرت) (الدوله | دول) العثمانيه (عشر | 10) (يناير | ديسمبر | نوفمبر | اكتوبر | راير | مارس | ابريل | مايو | يوليو | يونيو | سبتمبر) (ألف و أربعمئة و خمس و سبعون | 1475)

4) D 0034 AR AR اذكر دولة غزاها الالمان عام 1941؟--->

D AR AR 0034 (اذكر | اعرف | اسمع | اظن | اجد | اقول | اف | اتذكر | احسب | اتوقع | اكتب) دوله (غزاها | احتل | استعاد | حاصر | دمر | غزا | خرب | اخضع | استوطن | استرد | دخلت) الالمان (عام | سن) (ألف و تسعمئة و واحد و أربعون | 1941)

4.1.6.4. Queries transformation for Skip-Gram:

1) D 0008 AR AR كم عدد موتى اعصار 18 سبتمبر 1906 بهونج كونج؟ --->

D AR AR 0008 عدد موتي (اعصار | زلزال) (ثماني عشرة | 18) سبتمبر (ألف و تسعمئة و ست | 1906) بهونج كونج

2) D 0020 AR AR على اي ارتفاع انفجرت الطائرة الهندية يوم 23 يونيو 1985؟ --->

D AR AR 0020 ارتفاع (انفجرت | تحطمت | اصطدمت | اخترقت | انفجر | قصفت) الطائره الهنديه يوم (ثلاث و عشرون | 23) (يونيو | ديسمبر | مايو | اكتوبر | نوفمبر | يوليو | يناير | سبتمبر | ابريل | مارس | راير) (ألف و تسعمئة و خمس و ثمانون | 1985)

3) D 0028 AR AR ما هي الدولة التي هزمت الدولة العثمانية في 10 يناير 1475؟--->

D AR AR 0028 (الدوله | دول) (هزمت | هزم | خسرت | غزت | انتصرت) (الدوله | دول) العثمانيه (عشر | 10) (يناير | ديسمبر | نوفمبر | اكتوبر | راير | مارس | ابريل | مايو | يوليو | يونيو | سبتمبر) (ألف و أربعمئة و خمس و سبعون | 1475)

4) D 0034 AR AR اذكر دولة غزاها الالمان عام 1941؟--->
D AR AR 0034 (اذكر | اعرف | اسمع | اظن | اجد | اقول | اف | اتذكر | احسب | اتوقع | اكتب) دوله (غزاها | احتل | استعاد | حاصر | دمر | غزا | خرب | اخضع | استوطن | استرد | دخلت) الالمان (عام | سن) (ألف و تسعمئة و واحد و أربعون | 1941)

However, from the results found there is one relevant answer for one question, But the result extracted for the number not for the letters which is:
D AR AR 0008 عدد موتي (اعصار | زلزال) (ثماني عشرة | 18) سبتمبر (ألف و تسعمئة و ست | 1906) بهونج كونج

And this result was found because the majority of existing research write numbers as normal as is known and not in letters.

Finally, we can say there isn't an improvement for (CBOW and Skip-Gram models).

Notes:

1)Every time we specify big values for the interval we get few similar words, and every time we specify small values for the interval we get more similar words.

2) Most result obtained from google has a limit of 5 to 7 answers no more because of the large number of similar words given, since most questions have 11 similar words for each question word, and because of "+"sign.

4.1.6.5. Suggestions:

- It's better to write dates and numbers in letters in articles.

4.1.7. Test 07:

In this test we use the two word2vec models each time, and each models alone, in addition to the normalization, remove stop words, fix between quotation mark, fix Named Entity, transformation str to int (transformation of except the circumstance by using of roman numbers and leave number and the dates in the original case), extract similar words, stemming, replace "+" by " ", and raise the threshold of similarity, to take just similarities between this last and 1.0: (value similarity for central words= 0.7 and value similarity for Prepherical words = 0.7), We have set values largest than values of the test 06 for the threshold to get most similar words that have the same meaning with the origin word.

4.1.7.1. Results:

For the CBOW model, there is one query which has 4/10 relevant answers. And for the Skip-Gram model, there is one query which has 3/10 relevant answers.

4.1.7.2. Discussion:

We have set the interval to get similar words that have the same meaning with the origin word; the current results indicate that there is a change in 1 question of 50 questions in

total and with the use of word2vec Models (one time CBOW and other time Skip-Gram). This change is the transformation of numbers into letters, these questions for CBOW and Skip-Gram models are:

4.1.7.3. Queries transformation for CBOW model:

1) D0036 AR AR متى توفي البابا سلستين الثالث؟ --->

D AR AR 0036 (توفي | مات | اغتيل | توف) (البابا | بطرير | يوس | ندكت | بابا) سلستين III

4.1.7.4. Queries transformation for Skip-Gram:

1) D0036 AR AR متى توفي البابا سلستين الثالث؟

<--- When did Pope Celestine III die?

D AR AR 0036 (توفي | توف | مات | توفيت) (البابا | يوس | ندكت | بابا | اوربانوس | اينوسنت | انوسنت | شنود | نديكتوس | سيكتوس | يرلس) سلستين III

And this result was found because the majority of existing research write numbers as normal as is known and not in roman numbers.

Finally, As a result we can say there isn't an improvement for (CBOW and Skip-Gram models).

Notes:

1) Every time we specify big values for the interval we get few similar words, and every time we specify small values for the interval we get more similar words.

2) Most result obtained from google has a limit of 5 to 7 answers no more because of the large number of similar words given, since most questions have 11 similar words for each question word, and because of "+"sign.

4.1.7.5. Suggestions:

- It's better to transform into normal numbers.

4. Evaluation of results:

Here a table (Table 20) contains all type of reformulation with number of question reformulated for each type:

		Reformulation number										
	/	0	1	2	3	4	5	6	7	8	9	10
Test1(Question number)	/	6	4	7	3	5	5	3	2	4	6	5
Test2(Question number)	CBOW	18	3	3	5	7	9	2	1	1	1	0
	Skip-Gram	10	6	9	11	3	9	1	0	1	0	0
Test3(Question number)	CBOW	10	8	3	2	4	3	6	5	2	3	4
	Skip-Gram	6	5	3	4	5	4	6	7	5	2	3
Test4(Question number)	CBOW	2	3	5	3	5	4	4	3	3	7	11
	Skip-Gram	2	2	1	4	4	1	3	5	2	7	19
Test5(Question number)	CBOW	3	2	5	0	4	2	3	3	5	5	17
	Skip-Gram	1	3	2	4	2	3	3	6	3	8	15

Table 20: Number of questions reformulated by each type.

To show the efficiency of our work, we performed 7 tests. In most tests we used the word2vec algorithm with these two models (CBOW and Skip-Gram).

Tests	CBOW Results	Skip-Gram Results
Test02	26%	26%
Test03	42%	47%
Test04	62%	71%
Test05	67%	69%
Test06	Question has 1/10 of relevant answers.	Question that has 1/10 of relevant answers.
Test07	Question has 4/10 of relevant answers.	Question has 3/10 of relevant answers.

Table 21: Table of final results.

The first test, is to find the answers for the initial questions, without preprocessing and without using word2vec, after a relevance test we find the relevant percentage of answers that equals 47% for all questions.

All the tests that remain, they concern to find the answers for the initial questions, after a pretreatment and with use of word2vec and these two models, after the extraction of the answers a test of relevance is made.

As a conclusion, after the presentation of the results we see that we have an improvement just in two tests which have values of percentage higher than the values of the first test percentage.

For the first test in which there is an improvement, the value of percentage for Skip Gram model (71%>47%) it higher from the value of initial test (test01), and the value of percentage for CBOW model (62% > 47%) which is also higher from the value of initial test (test01)

For the second test in which there is an improvement, the value of percentage for Skip Gram model (69%>47%) it higher from the value of initial test (test01), and the value of percentage for CBOW model (67% > 47%) which is also higher from the value of initial test (test01)

5. Presentation of the application:

This application is just for well presented the reformulation of a query.

This application is developed by the python language using the HTML environment, with CSS and Java script for the design, also django for the frame works.

5.1.HTML (HyperText Markup Language):

First developed by Tim Berners-Lee in 1990, HTML is short for HyperText Markup Language. HTML is used to create electronic documents (called pages) that are displayed on the World Wide Web. Each page contains a series of connections to other pages called hyperlinks. Every web page you see on the Internet is written using one version of HTML code or another. [48]

5.2.CSS(Cascading Style Sheet):

CSS is the language for describing the presentation of Web pages, including colors, layout, and fonts. It allows one to adapt the presentation to different types of devices, such as large screens, small screens, or printers. CSS is independent of HTML and can be used with any XML-based markup language. The separation of HTML from CSS makes it easier to maintain sites, share style sheets across pages, and tailor pages to different environments. [49]

5.3.JavaScript:

Originally developed by Brendan Eich and originally known as LiveScript, the programming language JavaScript was renamed in 1995. JavaScript is an interpreted client-side scripting language that allows a web designer the ability to insert code into their web page. JavaScript is commonly placed into an HTML or ASP file, and runs directly from the web page. It is utilized to perform tasks such as printing the time and date, create a calendar, or other tasks that are not possible through plain HTML. [50]

5.4. Django:

Django is a high-level Python web framework that enables rapid development of secure and maintainable websites. Built by experienced developers, Django takes care of much of the hassle of web development, so you can focus on writing your app without needing to reinvent the wheel. It is free and open source, has a thriving and active community, great documentation, and many options for free and paid-for support. [51]

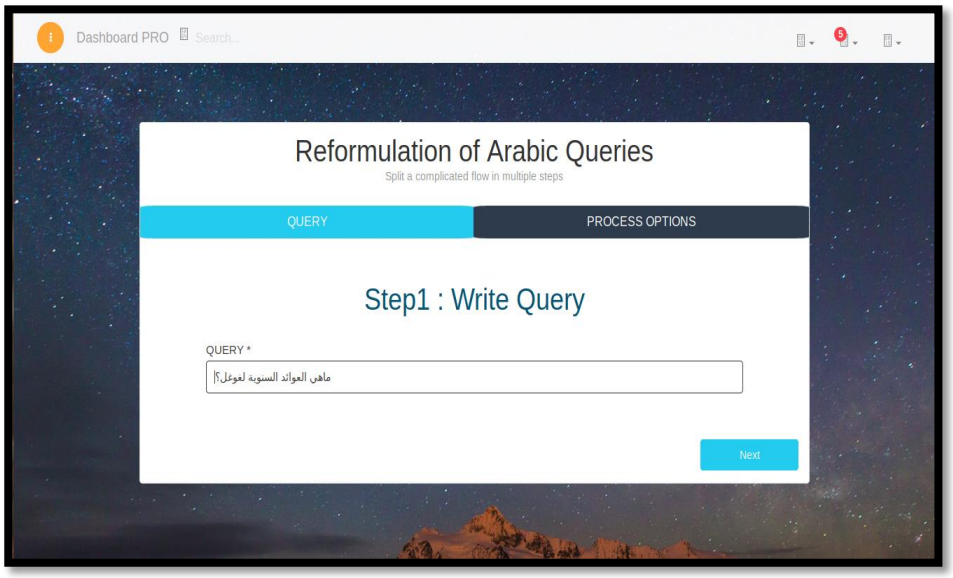


Figure 50: Query writing window.

The user writes his query in the "Query" box and click on the "NEXT" button.

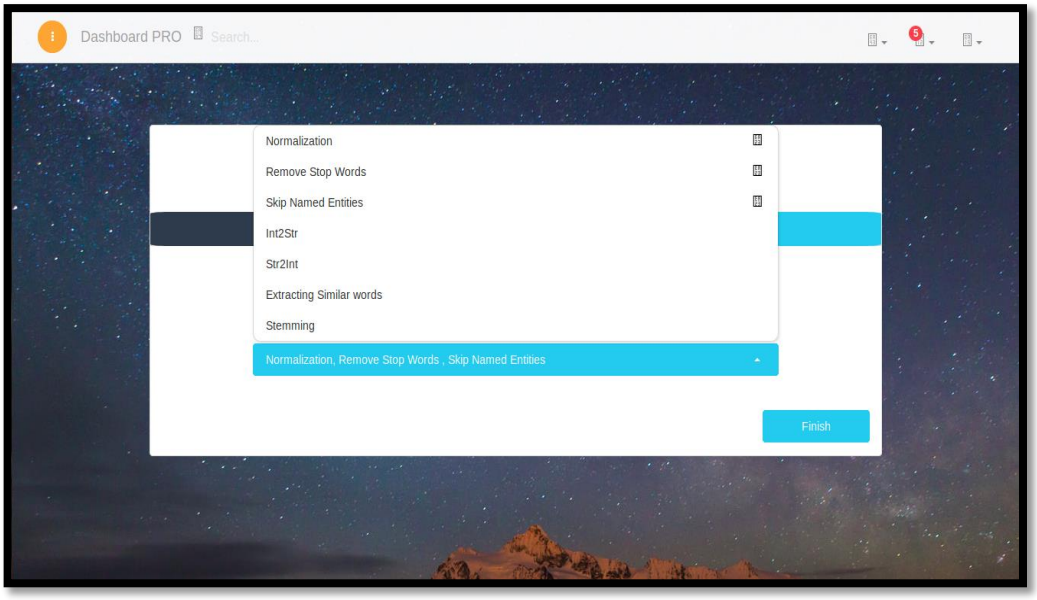


Figure 51: Window of choice the steps of Questions reformulation process used.

The user choice the steps of Questions reformulation process used, then click on the "Finish" button to see the results.

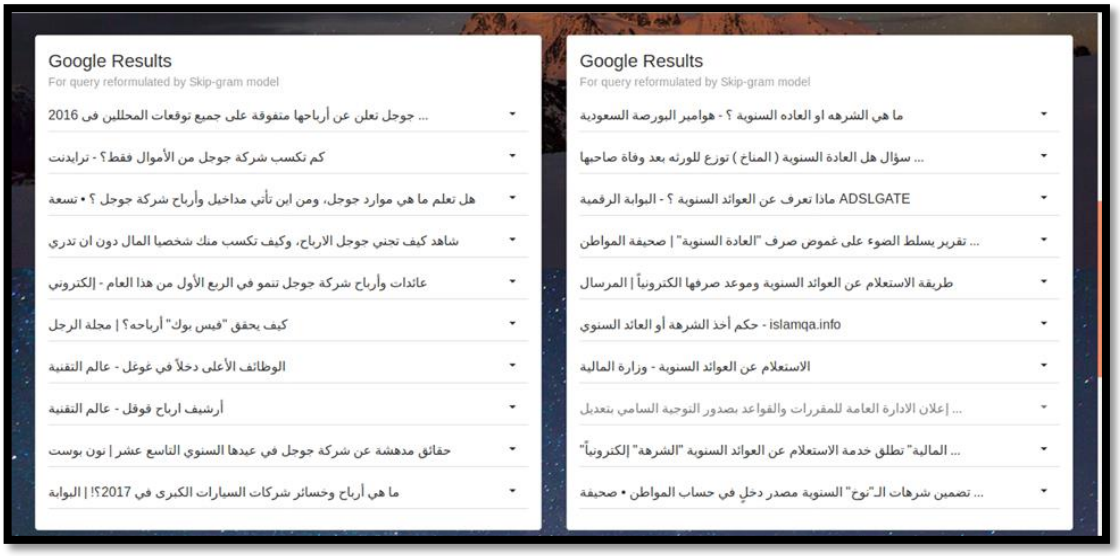


Figure 52: window of result.

6. Conclusion:

In this chapter, we have presented the implementation of our system Application of the Word2Vec algorithm to improve queries of Arabic question answering systems.

This system has the role of improving the questions presented in the dataset.

The results of our system were related to the choice of the techniques used that we proposed in our approach, it is essential factors, which participated in the improvement of the

performance of our system. And these most important steps: question reformulation process with their steps, the word2vec algorithm.

During the evaluation of our system, it was possible to make 7 tests, by the use of the pretreatment tools, and the use of the word2vec algorithm with its two models in each experiment as a main tool which our theme.

The evaluation also showed that our proposed approach can make an improvement (the goal).

In our implementation there is an improvement which is appear in the result of tests: 4 and 5, for the test 4 we have a percentage of (71%) for Skip Gram model, and of (62%) for CBOW model.

And for the test 5 we have a percentage of (67%) for Skip Gram model, and of (69%) for CBOW model.

General conclusion and perspectives

In our work, we have combined the use of two domains that are question Answering system, which aims to generate an answer to a question, gave with the word2vec algorithm that aims to extract similar words for a given word or phrase.

Our work aims to improve the questions that exist in Benjiba dataset which is choosed to make test. And more specifically questions in Arabic. These questions contain several different domains.

The approach that we proposed is based on the extraction of the relevant answers, related to the subject of given question, these answers are extracted by the use of Google API after we calculates the percentage of these answers.

We based on the use of percentages to see if there is an improvement between the initial and reformulated questions.

Our proposal encountered some difficulties with the syntactic complexity of the Arabic language, which posed a problem in the level of analysis of the questions, for that we apply pretreatment on the questions to eliminate some problems related to the Arabic language. With the use of word2vec to extract similar ones for words that remains after preprocessing, this is to find more accurate answers and related questions.

The results are encouraging but we can improve more, and as perspectives, we plan to complete the work for all questions. And also we consider the automatic recognition of central words to not make a list manually each time. Plus we also plan to create a model have a best quality, in addition, we plan to extract relevant answres automatic, then, we plan to update answers of Benajiba Dataset, finally, also we plan to take some Stop Words like words: كم من →عدد ,our work is limited in the use of deep learning because we haven't the opportunity to create a learning model because of some problem.

References

[Abdulla et al, 2013]

Nawaf A Abdulla, Nizar A Ahmed, Mohammed A Shehab, Mahmoud Al-Ayyoub, 2013, Arabic sentiment analysis: Lexicon-based and corpus-based, in: Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on, IEEE, Amman, Jordan, pp. 1_6, DOI: 10.1109/AEECT.2013.6716448.

[Abu Bakr Mohammad et al, 2017]

Soliman Abu Bakr Mohammad, El-Beltagy Samhaa R, EissaKareem, (2017, November 5-6), AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP, Center for Informatics Science, Nile University, Giza 12588, Egypt, DOI: 10.1016/j.procs.2017.10.117.

[Al-Jawfi, 2009]

Rashad Al-Jawfi, 2009, Handwriting arabic character recognition lenet using neural network, The International Arab Journal of Information Technology, Vol. 6, No pp. 304_309.

[Alotaibi, 2015]

Saud Saleh Alotaibi, 2015, Sentiment Analysis in the Arabic Language Using Machine Learning, In partial fulfillment of the requirements,
For the Degree of Doctor of Philosophy, Colorado State University, Department of Computer Science, Fort Collins, Colorado.

[Altowayan et al, 2016]

A. Aziz. Altowayan, Lixin. Tao, 2016, Word embeddings for arabic sentiment analysis, in: Big Data (Big Data), 2016 IEEE International Conference on, IEEE, pp. 3820_3825.

[Alotaibi et al, 2004]

Y.A.Alotaibi, 2004, Spoken arabic digits recognizer using recurrent neural networks, in: Signal Processing and Information Technology, Proceedings of the Fourth IEEE International Symposium on, IEEE.

[Aly et al, 2013]

Mohamed Aly, Amir Atiya, 2013, Labr: A large scale arabic book reviews dataset, in: ACL (2), pp. 494_498.

[Amr El-Desoky et al, 2013]

Amr El-Desoky Mousa, Hong-Kwang. Jeff Kuo, Lidia Mangu, Hagen Soltau, 2013, Morpheme-based feature-rich language models using deep neural networks for lvcsr of Egyptian arabic, in: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 8435_8439.

[Australian Institute of Health and Welfare (AIHW) 2007]

Australian Institute of Health and Welfare (AIHW) 2007, A guide to data development, AIHW Cat. no. HWI 94, Canberra: AIHW.

[Carlos Amaral & Dominique Laurent, 2006]

Carlos Amaral, Dominique Laurent, 2006, Implementation of a QA system in a real context.

[Chip Huyen & Danijar Hafner, 2012]

Chip Huyen, Danijar Hafner, 2012, Lecture note 4: How to structure your model in TensorFlow, "CS 20SI: TensorFlow for Deep Learning Research" (cs20si.stanford.edu).

[Dahou et al, 2016]

Abdelghani Dahou, Shengwu Xiong, Junwei Zhou, Mohamed Houcine Haddoud, Pengfei Duan, 2016, Word embeddings and convolutional neural network for arabic sentiment classification, in: COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 2418_2427.

[Delakis, 2008]

Manolis Delakis, Christophe Garcia, 2008, text detection with convolutional neural networks, in: VISAPP (2), pp. 290_294.

[Deng et al, 2009]

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, 2009, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, CVPR 2009. IEEE Conference on, IEEE, pp. 248_255, DOI: 10.1109/CVPR.2009.5206848.

[Deselaers et al, 2009]

Thomas Deselaers, Saša Hasan, Oliver Bender, Hermann Ney, 2009, A deep learning approach to machine transliteration, in: Proceedings of the Fourth Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 233_241.

[Dong et al, 2011]

Tiansi Dong, Ulrich Furbach, Ingo Glockner, Bjorn Pelzer, July 16-22, 2011, A Natural Language Question Answering System as a Participant in Human Q&A Portals, IJCAI'11 Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume ,Volume Three, Pages 2430-2435. Barcelona, Catalonia, Spain, DOI:10.5591/978-1-57735-516-8/IJCAI11-405.

[Elman 1990]

Jeffrey L. Elman, (1990), Finding structure in time, Cognitive science 14, pp. 179_211, University of California, San Diego.

[Elrazzaz et al, 2017]

Mohammed Elrazzaz, Shady Elbassuoni, Khaled Shaban, Chadi Helwe, 2017, Methodical evaluation of arabic word embeddings, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vol. 2, pp. 454_458.

[ElSahar et al, 2015]

Hady ElSahar, Samhaa R. El-Beltagy, 2015, Building large arabic multi-domain resources for sentiment analysis, in: CICLing (2), pp. 23_34.

[Graves et al, 2007]

Alex Graves, Santiago Fernandez, Juergen J. Schmidhuber, 2007, Multi-dimensional recurrent neural networks, in: ICANN (1), pp. 549_558.

[Graves et al, 2009]

Alex Graves, Jurgen Schmidhuber, 2009, Offline handwriting recognition with multidimensional recurrent neural networks, in: Advances in neural information processing systems, pp. 545_552.

[Graves, 2012]

Alex Graves, 2012, Offline arabic handwriting recognition with multidimensional recurrent neural networks, in: Guide to OCR for Arabic scripts, Springer, pp. 297_313.

[Girish]

K. GIRISH KUMAR, Word Embedding Techniques (word2vec, GloVe), Natural Language Processing Lab, Texas A&M University, Reading Group Presentation, last seen in 28 February 2018 Retrieved from http://nlp.cs.tamu.edu/resources/wordvectors.ppt.

[Gomez et al, 2016]

Lluis Gomez, Dimosthenis Karatzas, 2016, A fine-grained approach to scene text script iden-tification, in: Document Analysis Systems (DAS), 2016 12th IAPR Work- shop on, IEEE, pp. 192_197.

[Gomez et al, 2017]

Lluis Gomez, Anguelos Nicolaou, Dimosthenis Karatzas, 2017, Improving patch-based scene text script identif-cation with ensembles of conjoined networks, Pattern Recognition 67, pp. 85_96.

[Hani Al-Chalabi, Santosh Ray, Khaled Shaalan, 2015]

Hani Al-Chalabi, Santosh Ray, Khaled Shaalan, 2015, Semantic based Query Expansion for Arabic Question Answering Systems, 2015 First International Conference on Arabic Computational Linguistics, 978-1-4673-9155-9/15 $31.00 © 2015 IEEE, DOI 10.1109/ACLing.2015.25.

[Harrat et al, 2017]

Salima Harrat, Karima Meftouh, Kamel Smaïli, 2017, Machine translation for Arabic dialects (survey), Information Processing and Management, Elsevier.

[Jindal et al, 2017]

Vasu Jindal, 2017, A deep learning approach for arabic caption generation using roots-words, in: AAAI, pp. 4941_4942.

[Jindal et al, 2016]

Vasu Jindal, 2016, A personalized markov clustering and deep learning approach for arabic text categorization, ACL 2016, pp.145.

[KHUDHUR, 2014]

ABDULRAHMAN AHMED KHUDHUR, 2014, DIACRITICS OF ARABIC NATURAL LANGUAGE PROCESSING AND ITS QUALITY ASSESSMENT, A thesis submitted in partial fulfillment of the requirements for the award of Master of Computer Science (Software Engineering), The Department of Software Engineering, Faculty of Computer Science and Information Technology, University Tun Hussein Onn Malaysia.

[Kim et al, 2016]

Yoon Kim, Yacine Jernite, David Sontag, Alexander M. Rush, 2016, Character-aware neural language models, in: AAAI, pp. 2741_2749.

[Kingsbury et al, 2012]

Brian Kingsbury, Tara N. Sainath, Hagen Soltau, 2012, Scalable minimum bayes risk training of deep neural network acoustic models using distributed hessian-free optimization, in: Interspeech, pp. 10_13.

[LeCun et al, 1990]

Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, R. E. Howard, Wayne E. Hubbard, Lawrence D. Jackel,1990, Handwritten digit recognition with a back- propagation network, in: Advances in neural information processing systems , pp. 396_404.

[Lehmann, 2016]

Jens Lehmann, 2016, Question Answering: Applications and Challenges.

[Lahsen Abouenour, Karim Bouzoubaa, Paolo Rosso. December 2008]

Lahsen Abouenour, Karim Bouzoubaa, Paolo Rosso, December 2008, IMPROVING Q/A USING ARABIC WORDNET, Conference: Conference: Arabic Conference on Information Technology (ACIT), At Tunisia.

[Magnini, 2005]

Bernardo Magnini, 2005, Open Domain Question Answering: Techniques, Resources and Systems.

[Mahmoud Al-Ayyoub et al, 2017]

Mahmoud Al-Ayyoub, AyaNuseir, KholoudAlsmearat, YaserJararweh, Brij Gupta, (2017), Deep learning for Arabic NLP: A survey, *[CDATA[Journal of Computational Science]*, DOI: https://doi.org/10.1016/j.jocs.2017.11.011.

[Marwan Torki, Maram Hasanain, Tamer Elsayed, 2017]
Marwan Torki, Maram Hasanain, Tamer Elsayed, August 3 – 4,2017, QU-BIGIR at SemEval 2017 Task 3: Using Similarity Features for Arabic Community Question Answering Forums, Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), c 2017 Association for Computational Linguistics, Vancouver, Canada, pages 360–364.

[Mikolov et al, 2013]

Tomas Mikolov, Scott Wen-tau Yih, Geoffrey Zweig, 2013, Linguistic regularities in continuous space word representations, in: hlt-Naacl, Vol. 13, pp. 746_751.

[Mikolov et al, 2013]

Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, 7 September 2013, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

[Mikolov et al, 2013]

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, 2013, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, pp. 3111_3119.

[Mohit et al, 2017]
Mohit Jain, Minesh Mathew, C.V.Jawahar, 2017, Unconstrained scene text and video text recognition for arabic script, in: Arabic Script Analysis and Recognition (ASAR), 1st International Workshop on, IEEE, pp. 26_30.

[Momtazi, 2011]

Saeedeh Momtazi, 2011, Question Answering Systems, IT Systems Engineering University Potsdam.

[Mona Diab & Nizar Habash, 2014]
Mona Diab, Nizar Habash, 2014, Natural Language Processing of Arabic and its Dialects, EMNLP Doha Qatar Tutorial.
[Nabil et al, 2015]
Mahmoud Nabil, Mohamed Aly, Amir F Atiya, 2015, Astd: Arabic sentiment tweets dataset, in: EMNLP, pp. 2515_2519.

[Nakov et al, 2017]

Preslav Nakov, Doris Hoogeveen, Lluıs Marquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, KarinVerspoor, 2017,Semeval-2017 task 3: Community question answering, in: Proceedings of the 11th International Workshop on Semantic Evalua1tion (SemEval-2017), pp. 27_48.

[Neves, 2014]

Dr. Mariana Neves, 2014, Question Answering, IT Systems Engineering University Potsdam.

[Pennington et al, 2014]

Jeffrey Pennington, Richard Socher, Christopher D.Manning, October 25-29, 2014, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, pp. 1532_1543.

[Rashid et al, 2013]

Sheikh Faisal Rashid, Marc-Peter Schambach, Jörg Rottland, Stephan von der Null, 2013, Low resolution arabic recognition with multidimensional recurrent neural networks, in: Proceedings of the 4th International Workshop on Multilingual OCR, ACM, pp. 6, DOI:10.1145/2505377.2505385.

[Refaee et al, 2014]

Eshrag Refaee, Verena Rieser, 2014, An arabic twitter corpus for subjectivity and sentiment analysis, in: LREC, pp. 2268_2273.

[Romeo et al, 2017]

Salvatore Romeo, Giovanni Da San Martino, Yonatan Belinkov, Barron-Cedeno Alberto, Eldesouki Mohamed, Darwish Kareem, Hamdy Mubarak, Glass James, Moschitti Alessandro, (2017, August 15), Language Processing and Learning Models for Community Question Answering in Arabic, Information Processing & Management, DOI: 10.1016/j.ipm.2017.07.003.

[Sharma et al, 2015]

Nabin Sharma, Ranju Mandal, Rabi Sharma, UmapadaPal, Michael Blumenstein, 2015, Icdar2015 competition on video script identification (cvsi 2015), in: Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, IEEE, Tunis, Tunisia, pp. 1196_1200, DOI: 10.1109/ICDAR.2015.7333950.

[Shekarpour, 2015]

Saeedeh Shekarpour, 2015, A Tutorial on Question Answering Systems.

[Shiqing Hu, 2013]

Shiqing Hu Nijmegen, 2013, Dynamic QA Systems Using Knowledgebase Using an Improved Accuracy Method Based on How Net, A thesis submitted in fulfillment of the requirements for the degree of Master of Science, Information Science, Radboud Universiteit Nijmegen.

[Slimane et al, 2009]

Fouad Slimane, Rolf Ingold, Slim Kanoun, Adel M. Alimi, Jean Hennebert, 2009, A new arabic printed text image database and evaluation protocols, in: Document Analysis and Recognition, ICDAR'09. 10th International Conference on, IEEE, pp. 946_950, DOI: 10.1109/ICDAR.2009.155.

[Valentino Zocca et al, 2017]

Valentino Zocca, Spacagna Gianmario, Slater Daniel, Roelants Peter.(2017),*Python Deep Learning, Next generation techniques to revolutionize computer vision, AI, speech and data analysis.*

[Wei Kor, 2005]

Kian Wei Kor, 2005**,** Improving Answer Precision and Recall of List Questions, Master of Science, School of Informatics, University of Edinburgh.

[Xin Rong, 2015]

Xin Rong, (2015, November 7), Word Embedding Explained and Visualized. Retrieved from https://docs.google.com/presentation/d/1yQWN1CDWLzxGeIAvnGgDsIJr5xmy4dB0VmHFKkLiibo/pub?start=false&loop=false&delayms=3000&slide=id.ge79682746_0_132.

[Yonatan Belinkov, Mitra Mohtarami, Scott Cyphers, James Glass, 2015]

Yonatan Belinkov, Mitra Mohtarami, Scott Cyphers, James Glass, June 4-5, 2015, VectorSLU: A Continuous Word Vector Approach to Answer Selection in Community Question Answering Systems, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, c 2015 Association for Computational Linguistics, pp. 282–287.

[Yousfi et al, 2014]

Sonia Yousfi, Sid-Ahmed Berrani, Christophe Garcia, 2014, Arabic text detection in videos using neural and boosting-based approaches: Application to video indexing, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, pp. 3028_3032.

[Yousfi et al, 2015]

Sonia Yousfi, Sid-Ahmed Berrani, Christophe Garcia, 2015, Alif: A dataset for arabic embedded text recognition in tv broadcast, in: Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, IEEE, pp. 1221_1225.

[Yousfi et al, 2015]

Sonia Yousfi, Sid-Ahmed Berrani, Christophe Garcia, 2015, Deep learning and recurrent connectionist-based approaches for arabic text recognition in videos, in: Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, IEEE, pp. 1026_1030.

[Yousfi, 2017]

Sonia Yousfi, Sid-Ahmed Berrani, Christophe Garcia, 2017, Contribution of recurrent connectionist language models in improving lstm-based arabic text recognition in videos, Pattern Recognition 64, pp. 245_254.

Web references

[1]: Savan Visalpara, S. (2017, March 2). What is deep learning? Why is this a growing trend in machine learning? Why not use SVMs?. Retrieved from https://www.quora.com/What-is-deep-learning-Why-is-this-a-growing-trend-in-machine-learning-Why-not-use-SVMs(Poster)

[2]: STFALCON. (2017, December 13).Deep Learning: Definition, Benefits, and Challenges. Retrieved from https://www.upwork.com/hiring/for-clients/deep-learning-definition-benefits-challenges/ (Article)

[3]: NAVDEEP SINGH GILL. (2017, July 6). Log Analytics With Deep Learning And Machine Learning. Retrieved from https://www.upwork.com/hiring/for-clients/log-analytics-deep-learning-machine-learning/ (Article)

[4]: M. Tim Jones, M. (2017, August 17). Recurrent neural networks deep dive. Retrieved from https://www.ibm.com/developerworks/library/cc-cognitive-recurrent-neural-networks/index.html (Article)

[5]: Ankit Gupta, A. (2017, November 8). What is Deep Learning and Neural Network. Retrieved from http://www.thewindowsclub.com/deep-learning-and-neural-network (Article)

[6]: Aditya Thakker, A. (2017, Jun 18). Introduction to Word2Vec & How it Works. Retrieved from https://www.adityathakker.com/introduction-to-word2vec-how-it-works/(Poster)

[7]: Sanjeev Arora. (2015, December 12). Semantic Word Embeddings. Retrieved from http://www.offconvex.org/2015/12/12/word-embeddings-1/(Poster)

[8]: Zareen Farooqui, Z. (2016, July 22). Word2Vec on Harry Potter. Retrieved from https://medium.com/zareen-farooqui/word2vec-on-harry-potter-e3748ab27e9 (Article)

[9]: Julian Gilyadov, J. (2017, March 23).Word2Vec Explained. Retrieved from https://israelg99.github.io/2017-03-23-Word2Vec-Explained/ (Blog)

[10]: Gabriel Mordecki, G. (2017, December 7). Word embeddings: how to transform text into numbers. Retrieved from https://monkeylearn.com/blog/word-embeddings-transform-text-numbers/ (Blog)

[11]: Matthias, M. (2017, September 7). Semantic trees for training word embeddings with hierarchical softmax. Retrieved from https://blog.lateral.io/2017/09/semantic-trees-hierarchical-softmax/ (Poster)

[12]: Kejin Jin, K. Junpei Zhong, J. (2017, Jun 2). Word2vec: Google's New Leap Forward on the Vectorized Representation of Words. Retrieved from https://medium.com/@Synced/word2vec-googles-new-leap-forward-on-the-vectorized-representation-of-words-e8e505a31595 (Poster)

[13]: Artificial Intelligence Tutorial. (2018). Retrieved from https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language processing.htm (Tutorial)

[14]: Scott, R. (2017, December 20). What is Natural Language Processing?. Retrieved from http://stackabuse.com/what-is-natural-language-processing/(Course)

[15]: Matt, K. (2016, AUGUST 11).Introduction to Natural Language Processing (NLP). Retrieved from https://blog.algorithmia.com/introduction-natural-language-processing-nlp/ (Blog)

[16]: Examples of natural language processing systems in artificial intelligence. (2017). Retrieved from http://www.expertsystem.com/examples-natural-language-processing-systems-artificial-intelligence/ (Blog)

[17]: Question. (2018, April 10). Retrieved from https://en.wikipedia.org/wiki/Question (Poster)

[18]: Normalization. (2017, November 8). Retrieved from https://en.wikipedia.org/wiki/Normalization (Poster)

[19]: Google Search Help. (2018). Retrieved from https://support.google.com/websearch/answer/2466433?hl=en (HELP FORUM)

[20]: xpo6. (2009, April 14). List of English Stop Words. Retrieved from http://xpo6.com/list-of-english-stop-words/ (Blog)

[21]: Stop words. (2017, April 26). Retrieved from https://www.computerhope.com/jargon/s/stopword.htm (Poster)

[22]: Named entity. (2018, January 4). Retrieved from https://en.wikipedia.org/wiki/Named_entity (Poster)

[23]: Word2vec. (2018, February 26). Retrieved from https://en.wikipedia.org/wiki/Word2vec (Poster)

[24]: R, Margaret. (2018, January). stemming. Retrieved from https://searchenterpriseai.techtarget.com/definition/stemming (Poster)

[25]: Stemming. (2018, April 7). Retrieved from https://en.wikipedia.org/wiki/Stemming (Poster)

[26]: Information retrieval. (2018, April 24). Retrieved from https://en.wikipedia.org/wiki/Information_retrieval (Poster)

[27]: Advantages and Disadvantages of Python Programming Language. (2017, April 24). Retrieved from https://medium.com/@mindfiresolutions.usa/advantages-and-disadvantages-of-python-programming-language-fd0b394f2121 (Poster)

[28]: w3resource. (2018, April 14). Python Tutorial. Retrieved from https://www.w3resource.com/python/python-tutorial.php(Tutorial)

[29]: M, Prakhar.(2017, December 4). Word Embeddings with Gensim. Retrieved from https://becominghuman.ai/word-embeddings-with-gensim-68e6322afdca (Blog)

[30]: Esther Vaati . (2017, November 21). Introduction to Regular Expressions in Python. Retrieved from http://stackabuse.com/introduction-to-regular-expressions-in-python/ (tutorial)

[31]: Python Regular Expressions. (2012, December 13). Retrieved from https://developers.google.com/edu/python/regular-expressions (courses)

[32]: Aditi Gupta. (2018, June 16). Unicodedata – Unicode Database in Python. Retrieved from https://www.geeksforgeeks.org/unicodedata-unicode-database-python/ (Article)

[33]: 7.9. unicodedata — Unicode Database. (2018, April 29). Retrieved from https://docs.python.org/2/library/unicodedata.html (Page)

[34]: Taha Zerrouki (طه زروقي). (2018, May 5). An Arabic language library for Python. Retrieved from https://github.com/linuxscout/pyarabic(GitHub)

[35]: Python strings. (2018, February 26). Retrieved from http://zetcode.com/lang/python/strings/ (tutorial)

[36]: Google APIs. (2018, January 4). Retrieved from https://en.wikipedia.org/wiki/Google_APIs (Page)

[37]: Posts Tagged Google API. (2012, April 10). Retrieved from http://blog.soton.ac.uk/eventus/tag/google-api/ (Post)

[38]: Tim Hoffmann. (2018, May 26). Matplotlib. Retrieved from https://github.com/matplotlib/matplotlib (GitHub)

[39]: Taha Zerrouki (طه زروقي). (2018, April 27). Tashaphyne. Retrieved from https://github.com/linuxscout/tashaphyne (GitHub)

[40]: Brendan Dolan-Gavitt. (2016, December 18). PANDA. Retrieved from https://github.com/moyix/panda (GitHub)

[41]: NumPy. (2018). Retrieved from http://www.numpy.org/index.html (Page)

[42]: Loïc Estève. (2018, June 7). scikit-learn. Retrieved from https://github.com/scikit-learn/scikit-learn (GitHub)

[43]: K, darwish. (2016, October 29). Farasa. Retrieved from https://github.com/kdarwish/Farasa (GitHub)

[44]: Abu Bakr Soliman. (2018, April 01). AraVec2.0. Retrieved from https://github.com/bakrianoo/aravec/blob/master/README.md (GitHub)

[45]: Stop words. (2017, April 26). Retrieved from https://www.computerhope.com/jargon/s/stopword.htm (Poster)

[46]: Taha Zerrouki (طه زروقي) . (2018, May 3). Arabic Stop Word List. Retrieved from https://github.com/linuxscout/arabicstopwords (GitHub)

[47]: Taha Zerrouki (طه زروقي). (2018, May 5). Pyarabic. Retrieved from https://github.com/linuxscout/pyarabic (GitHub)

[48]: Computer Hope. (2017, December 20). HTML. Retrieved from https://www.computerhope.com/jargon/h/html.htm (Help Pages)

[49]: W3C. (2016). HTML & CSS. Retrieved from https://www.w3.org/standards/webdesign/ (Web Design and Applications)

[50]: Computer Hope. (2017, December 29). JavaScript**.** Retrieved from https://www.computerhope.com/jargon/j/javascri.htm (Help Pages)

[51]: warisali2. (2018, Februry 16). Django introduction. Retrieved from https://developer.mozilla.org/en-US/docs/Learn/Server-side/Django/Introduction(Article)

Annex

In this table we find all the questions of Benajiba Dataset:

Number	Question	Traduction
D 0001 AR AR	كم هي العوايد السنوية لغوغل؟	How many are the annual joys of Google?
D 0002 AR AR	كم ينفق على التدخين سنويا في الولايات المتحدة الامريكية؟	How much is spent on smoking annually in the United States?
D 0003 AR AR	كم تبلغ القيمة المادية لجائزة المغرب للكتاب؟	How much is the material value of the Morocco Book Award?
D 0004 AR AR	كم يبلغ طول ضلع بركة الحمام؟	How long is the rib bath pool?
D 0005 AR AR	كم يبلغ ارتفاع برج ايفل؟	How high is the Eiffel Tower?
D 0006 AR AR	كم كان عمر الامير سعود عبد العزيز بن متعب الرشيد عندما تولى الحكم؟	How old was Prince Saud bin Abdul Aziz bin Mutib al-Rashid when he took power?
D 0007 AR AR	في اي سن بدأ كات ستيفنز مسيرته الفنية؟	At what age did Cat Stevens start his career?
D 0008 AR AR	كم عدد موتى اعصار 18 سبتمبر 1906 بهونج كونج؟	How many died of typhoon September 18, 1906 in Hong Kong?
D 0009 AR AR	كم كانت قوة زلزال هوكايدو؟	How powerful was the Hokkaido earthquake?
D 0010 AR AR	كم كان عدد قتلى ومصابي زلزال هوكايدو؟	How many deaths were victims of the Hokkaido earthquake?
D 0011 AR AR	كم من كتاب الفه وجالينوس في الطب والصيدلة؟	How Much of Book write Galen in Medicine and Pharmacy?
D 0012 AR AR	كم تبلغ قوة انفجار قنبلة نووية؟	How powerful is a nuclear bomb?
D 0013 AR AR	ما هو معدل درجة الحرارة في فصل الصيف؟	What is the temperature in summer?
D 0014 AR AR	كم عدد الاعضاء الغير الدائمة في مجلس الامن؟	How many non-permanent members of the Security Council?
D 0015 AR AR	كم عدد اعضاء مجلس الامن؟	How many members of the Security Council?
D 0016 AR AR	كم عدد جنود الجيش الذي حرر عكا من يد الصليبيين تحت قيادة السلطان الاشرف خليل؟	How many army soldiers liberated Acre from the hands of the Crusaders under the leadership of Sultan Khalil?
D 0017 AR AR	ما هي المدة التي يستغرقها كوكب عطارد في دورة واحدة حول الشمس؟	How long does Mercury take in one cycle around the Sun?
D 0018 AR AR	ما هي المدة التي يستغرقها كوكب الارض في مداره حول الشمس؟	How long does the planet take in its orbit around the sun?
D 0019 AR AR	كم عدد انواع السلحفيات؟	How many species of turtles?
D 0020 AR AR	على اي ارتفاع انفجرت الطائرة الهندية يوم 23 يونيو 1985؟	At what altitude did the Indian plane explode on June 23, 1985?
D 0021 AR AR	كم من سفينة تجارية دمرت في حرب الناقلات؟	How much of a commercial ship destroyed in a tanker war?
D 0022 AR AR	كم من كيلوبايت في ميجابايت واحد؟	How many kilobytes per megabyte?
D 0023 AR AR	ما هو الوزن الجزيئي لغاز الميثان؟	What is the molecular weight of methane?
D 0024 AR AR	ما اسم البرنامج الاذاعي الاجتماعي الذي يقدمه الفنان فؤاد المهندس؟	What is the name of the social radio program presented by the artist Fouad Engineer?
D 0025 AR AR	اذكر احد ابطال فيلم"غصن الزيتون"؟	Remember one of the heroes of the film "olive branch"?
D 0026 AR AR	من هو مخرج فيلم"الملكة الافريقية"؟	Who is the director of "The African Queen"?
D 0027 AR AR	في اي دولة تولى الشعلان منصب وزير الدفاع؟	In what country did Shaalan assume the position of Minister of Defense?
D 0028 AR AR	ما هي الدولة التي هزمت الدولة العثمانية في 10 يناير 1475؟	What is the state that defeated the Ottoman Empire on 10 January 1475?
D 0029 AR AR	ما هي الشركة التي اشترت تايم وارنر؟	What is the company that bought Time Warner?
D 0030 AR AR	اين ولد ابن بطوطة؟	Where was Ibn Battuta born?
D 0031AR AR	اين ولد كليف ريتشارد؟	Where was Clive Richard born?
D 0032 AR AR	في اي ولاية توجد والا والا؟	In which state is there Walla Walla?
D 0033AR AR	في اي ولاية توجد بلدة ماتيس؟	In what state is Matisse located?
D 0034 AR AR	اذكر دولة غزاها الألمان عام 1941؟	Remember a country conquered by the Germans in 1941?
D 0035 AR AR	في اي دولة تولى سيرجي لافروف منصب وزير الخارجية؟	In which country did Sergei Lavrov assume the post of Foreign Minister?
D 0036 AR AR	متى توفي البابا سلستين الثالث؟	When did Pope Celestine III die?
D 0037 AR AR	متى استشهد عبد القادر الحسيني؟	When did Abdul Qadir al-Husseini die?
D 0038 AR AR	متى اصدرت لأول مرة نيويورك تايمز؟	When did the New York Times debut?

D 0039 AR AR	متى ولد المطرب كاظم الساهر؟	When was the singer Kazim El Saher born?
D 0040 AR AR	متى اصدرت اغنية"في مدرسة الحب"؟	When did the song "In the School of Love" be released?
D 0041AR AR	متى اسست الشركة الوطنية الفرنسية للسكك الحديدية؟	When was the French National Railway Company established?
D 0042 AR AR	متى اغتيل عبد العزيز الرنتيسي؟	When was Abdel Aziz Rantisi assassinated?
D 0043 AR AR	متى ولد طالباني؟	When was Talabani born?
D 0044 AR AR	متى بدأ لينين الثورة البلشيفية؟	When did Lenin begin the Bolshevik revolution?
D 0045 AR AR	متى اكتشف حجر رشيد ابان؟	When was Rosetta Stone discovered?
D 0046 AR AR	متى انضمت سلطنة عمان الى جامعة الدول العربية؟	When did Oman join the League of Arab States?
D 0047 AR AR	ما هي الرياضة التي يمارسها رونالدو؟	What is Ronaldo's sport?
D 0048 AR AR	ما هي جنسية عباس الموسوي؟	What is the nationality of Abbas al-Musawi?
D 0049 AR AR	في اي معركة قتل تشارلز الاصلع؟	In what battle did Charles the bald man die?
D 0050 AR AR	في اي معركة هزم الفرنسيون براند نبرغ؟	In which battle did the French defeated Brand Nürburg?
D 0051 AR AR	من هو الملك الفرنسي الذي وقع التحالف العسكري الامريكي الفرنسي عام 1778؟	Who is the French king who signed the Franco-American military alliance in 1778?
D 0052 AR AR	في اي سنة توفي محمد بن ادريس الشافعي؟	In what year did Muhammad bin Idris al-Shafei die?
D 0053 AR AR	ما مكافأة محمد علي للقبائل التي هزمت عربان غزة؟	What rewarded Muhammad Ali to the tribes that defeated the Arabs of Gaza?
D 0054 AR AR	ما هو التاريخ الرسمي لاتحاد المانيا؟	What is the official history of the German Federation?
D 0055 AR AR	كيف يسمى الخط الذي يرسم الحدود بين المانيا وبولندا؟	How is the line drawing the border between Germany and Poland called?
D 0056 AR AR	ما هي المعاهدة التي أقامت الحد للحرب بين اليمن والسعودية؟	What treaty established the limit for the war between Yemen and Saudi Arabia?
D 0057 AR AR	عن اي دولة استقلت تيمور الشرقية؟	From which country do East Timor take their independent?
D 0058 AR AR	ما هي وحدة قياس الزلزال؟	What is the earthquake measuring unit?
D 0059 AR AR	الى من تنسب وحدة القياس ريختر؟	To whom is the Richter unit attributed?
D 0060 AR AR	في اي يوم انتخب جورج وانشطن لرئاسة الولايات المتحدة الامريكية؟	On which day was George W. Bush elected to the presidency of the United States of America?
D 0061 AR AR	ما هو الاسم الذي اتخذته يوغوسلافيا عندما غيرته؟	What is the name taken by Yugoslavia when you changed it?
D 0062 AR AR	في اي دولة توجد مدينة غواتيمالا؟	In which country is Guatemala City located?
D 0063 AR AR	اي دولة احتلتها بريطانيا سنة 1840؟	Which country was occupied by Britain in 1840?
D 0064 AR AR	ما هو نوع أول حاسب مزود بفأرة؟	What is the first computer with a mouse?
D 0065 AR AR	ها هي المنظمة التي تعمل على تحديد معايير دولية؟	Here is the organization that sets international standards?
D 0066 AR AR	الى اي دولة سلم باكيتو؟	To which country did Pacito concede?
D 0067 AR AR	من أسس منظمة"مركز الخدمات"؟	Who founded the Service Center?
D 0068 AR AR	من أسس منظمة"مؤسسة المسجد الاسلامي"؟	Who founded the Islamic Masjid Foundation?
D 0069 AR AR	ما هي المنظمة التي عقدت اتفاقية اوسلو 1994؟	What is the organization that held the Oslo Accords 1994?
D 0070 AR AR	في اي دولة أقام المشير انقلابا عسكريا؟	In what country did Marshal set up a military coup?
D 0071 AR AR	ما هو المجلس الذي يترأسه عبد الرحمن سوار الذهب المشير؟	What is the council headed by Abdul Rahman al-Zahar al-Mashir?
D 0072 AR AR	في اي دولة أقيمت الالعاب الاولمبية الثالثة؟	In which country was the third Olympic Games held?
D 0073 AR AR	من اي اسم اشتق اسم مدينة قورنيا؟	From what name did the name of the city of Qurnia derive?
D 0074 AR AR	في اي قرن بني مسجد الحبشي؟	In what century was the mosque of Abyssinia built?
D 0075 AR AR	في اي دولة توجد ولاية البحر الأحمر؟	In which state is the Red Sea State?
D 0076 AR AR	ما هي الفرقة التي قادها رومل؟	What band did Rommel lead?
D 0077 AR AR	أين عاش شعب بوش؟	Where did the people of Bush live?
D 0078 AR AR	في اي بطولة عالمية شارك نادي النصر السعودي؟	In which world championship participated Al-Nasr Saudi Club?

D 0079 AR AR	ما هو الفريق الذي واجه فريق البرازيل في أول مبارة عالمية له؟	What was the team that faced Brazil in their first World Cup match?
D 0080 AR AR	ما هي جنسية المدرب العالمي كارلوس البرتو؟	What is the nationality of the international coach Carlos Alberto?
D 0081 AR AR	أين ولد مارك فان بوميل؟	Where was Mark van Bommel born?
D 0082 AR AR	أين أقيم كأس العالم لكرة القدم سنة 1986؟	Where was the FIFA World Cup held in 1986?
D 0083 AR AR	في أي سنة وصل فريق صربيا والجبل الاسود الى نهائيات كأس العالم لكرة القدم لأول مرة؟	In which year did the Serbia and Montenegro team reach the World Cup finals for the first time?
D 0084 AR AR	من أسس الدولة الرستمية؟	From the foundations of the Rustamian state?
D 0085 AR AR	كيف يسمى أبريل في بلاد الشام؟	How to call April in the Levant?
D 0086 AR AR	بكم من كيلومتر تبعد بني وليد عن طرابلس؟	Are you miles away from Bani Walid from Tripoli?
D 0087 AR AR	في أي مدينة كان الغساسنة؟	In which city was Ghassanah?
D 0088 AR AR	في أي مدينة ولد علي بن حزم الاندلسي؟	In which city was Ali bin Hazm al-Andalusi born?
D 0089 AR AR	ما هي أعلى درجة حرارة سجلت في لندن؟	What is the highest temperature recorded in London?
D 0090 AR AR	ما هي اللغة الرئيسية في أندونيسيا؟	What is the main language in Indonesia?
D 0091 AR AR	متى فاز الفريق البريطاني لأول مرة بدوري أبطال أوروبا؟	When did the British team win for the first time in the Champions League?
D 0092 AR AR	ما هو المنصب الذي شغله شارل ديغول في فرنسا؟	What was Charles de Gaulle's job in France?
D 0093 AR AR	ما هي الشركة التي تملك تقنية انفراسيرتش؟	What is the company that owns Infrasearch technology?
D 0094 AR AR	في اي سنة استقل سلوبودان ميلوسوفيتش؟	In what year did Slobodan Milosevic take over?
D 0095 AR AR	ما هو الاسم المختصر لنينغشيا؟	What is the short name of Ningxia?
D 0096 AR AR	ما هي نسبة النمساويين الذين صوتوا لصالح انضمام النمسا الى الاتحاد الاوروبي؟	What percentage of Austrians voted for Austria's accession to the European Union?
D 0097 AR AR	من اخترع التلسكوب الانكساري؟	Who invented the refractive telescope?
D 0098 AR AR	من ألف كتاب الصيدلة"بوتساو"؟	Who is the book of pharmacy "Potsau"?
D 0099 AR AR	ما هو الاسم الحقيقي لجينكز خان؟	What is the real name of Genghis Khan?
D 0100 AR AR	ما هي أول مستعمرة بريطانية في أمريكا الشمالية؟	What is the first British colony in North America?
D 0101 AR AR	في أي دولة تقع محافظة ايواتيه؟	In any state located Iwate Prefecture?
D 0102 AR AR	ما هي آخر القلاع البوذية التي سقطت في اليابان؟	What are the last Buddhist castles that fell in Japan?
D 0103 AR AR	في اي قارة تقع ماليزيا؟	In which continent is Malaysia located?
D 0104 AR AR	في أي دولة توجد محافظة باتاني؟	In any country is the province of Pattani?
D 0105 AR AR	في اي سنة اتحدت ماليزيا؟	In what year did Malaysia unite?
D 0106 AR AR	ما معنى بي أو سي؟	What is the meaning of BOC?
D 0107 AR AR	في أي مدينة ولد عبد العزيز جاويش؟	In which city was Abdul Aziz Jawish born?
D 0108 AR AR	في اي مدينة القى جورج واشنطن أول خطاب سنوي؟	In which city did George Washington give his first annual speech?
D 0109 AR AR	اين توجد مدينة جنين؟	Where is the city of Jenin?
D 0110 AR AR	أي دولة حكمها الملك فيصل؟	Which country is ruled by King Faisal?
D 0111 AR AR	في اي دولة توجد الأهرامات؟	In which country do pyramids exist?
D 0112 AR AR	أين توفي الامام الشافعي؟	Where did Imam Shafei die?
D 0113 AR AR	ما هي الدولة التي حكمها جورج الخامس؟	What is the state ruled by George V?
D 0114 AR AR	أين ولد عبد الرحمن سوار الذهب؟	Where was Abdul Rahman al-Zahab born?
D 0115 AR AR	في أي دولة تنشط حركة الوفاق الوطني العراقي؟	In which country is the Iraqi National Accord Movement active?
D 0116 AR AR	ما هو عنوان اول قصة قصيرة نشرت للكاتب محمد مستجاب؟	What is the title of the first short story published by the writer Mohammed Mustap?
D 0117 AR AR	متى هجمت إيطاليا على اثيوبيا؟	When did Italy attack Ethiopia?
D 0118 AR AR	متى سلم السوفييت الحكم لجمهورية المانيا؟	When did the Soviets hand over power to the Republic of Germany?
D 0119 AR AR	متى اعتقل الفرنسيون أحمد بومراق؟	When the French arrested Ahmed Boumerak?
D 0120 AR AR	متى ولد شون دو هوان (رئيس كوريا الجنوبية)؟	When was Shun Do Huan (President of South Korea)?
D 0121 AR AR	متى كانت اول سنة لحكم الملك زوسر؟	When was the first year of the rule of King Zoser?
D 0122 AR AR	متى استقلت موناكو عن جنوة؟	When Monaco resigned from Genoa?
D 0123 AR AR	من هو أبقراط أبو الطب؟	Who is Hippocrates of Medicine?

D 0124 AR AR	من هو نزار القباني؟	Who is Nizar Qabbani?
D 0125 AR AR	من هو محمد عبد الحليم عبد الله؟	Who is Mohamed Abdel Halim Abdullah?
D 0126 AR AR	من هو جين كيزنل؟	Who is Kevin Quinnel?
D 0127 AR AR	من هو حسين رياض؟	Who is Hussein Riad?
D 0128 AR AR	من هو نيكولايفتش تولستوي؟	Who is Nikolaevich Tolstoy?
D 0129 AR AR	من هو جول فيرن؟	Who is Jules Verne?
D 0130 AR AR	من هو ويلز؟	Who is Wales?
D 0131 AR AR	من هو الحسين بن طلال؟	Who is Hussein Bin Talal?
D 0132 AR AR	من هو علال الفاسي؟	Who is Alal Fassi?
D 0133 AR AR	من هو صدام حسين؟	Who is Saddam Hussein?
D 0134 AR AR	من هو ميغيل بريمو دي رفييرا؟	Who is Miguel Primo de Raviera?
D 0135 AR AR	من هو ماركو بولو؟	Who is Marco Polo?
D 0136 AR AR	من هو موشيه دايان؟	Who is Moshe Dayan?
D 0137 AR AR	من هو يوجيني لو دوك؟	Who is Eugenie Le Duc?
D 0138 AR AR	من هو بابرشاه الثاني؟	Who is Babarsha II?
D 0139 AR AR	من هو عمر الخيام؟	Who is the age of tents?
D 0140 AR AR	من هي اندير غاندي؟	Who is Indir Gandhi?
D 0141 AR AR	من هو حسن نصر الله؟	Who is Hassan Nasrallah?
D 0142 AR AR	من هو فالديمير بوتين؟	Who is Vladimir Putin?
D 0143 AR AR	من هو حسن البنا؟	Who is Hasan al-Banna?
D 0144 AR AR	من هو ادوارد السابع؟	Who is Edward VII?
D 0145 AR AR	من هو اخناتون أمنحبوت الرابع؟	Who is Akhenaton Amenhotot IV?
D 0146 AR AR	من هو مرنبتاح؟	Who is Merneptah?
D 0147 AR AR	من هي مارغريت الثانية؟	Who is Margaret II?
D 0148 AR AR	ما هي فورد؟	What is Ford?
D 0149 AR AR	ما هي هيئة الاذاعة البريطانية؟	What is the BBC?
D 0150 AR AR	ما هي سي ان ان؟	What is CNN?
D 0151 AR AR	ما هي الجارديان؟	What are the Guardian?
D 0152 AR AR	ما هي يو.اس.اي توداي؟	What is USA Today?
D 0153 AR AR	ما هو السي++؟	What is C ++?
D 0154 AR AR	ما هي اليونسكو؟	What is UNESCO?
D 0155 AR AR	ما هي أي بي إم؟	What is IBM?
D 0156 AR AR	ما هو أف سي أيندهوفن؟	What is FC Eindhoven?
D 0157 AR AR	ما هي فوربز؟	What is Forbes?
D 0158 AR AR	ما هي أوراكل؟	What is Oracle?
D 0159 AR AR	ما هي لوبريف؟	What is LOUBREF?
D 0160 AR AR	ما هي الأحمدية؟	What is Ahmadiyya?
D 0161 AR AR	ما هي صنداي تايمز؟	What are the Sunday Times?
D 0162 AR AR	ما هي الجزيرة؟	What is the island?
D 0163 AR AR	ما هي نيويورك تايمز؟	What is the New York Times?
D 0164 AR AR	ما هي فايننشال تايمز؟	What is the Financial Times?
D 0165 AR AR	ما هو حزب الاتحاد الاشتراكي للقوات الشعبية؟	What is the Socialist Union of Popular Forces?
D 0166 AR AR	ما هو الاتحاد الوطني الكردستاني؟	What is the Patriotic Union of Kurdistan?
D 0167 AR AR	ما هي سيليكون فالي؟	What is Silicon Valley?
D 0168 AR AR	ما هي صن ميكروسيستمز؟	What is Sun Microsystems?
D 0169 AR AR	ما هي سوني؟	What is Sony?
D 0170 AR AR	ما هي فيلبس؟	What is Phelps?
D 0171 AR AR	ما هي "هني ويل"؟	What is Honeywell?
D 0172 AR AR	من هو ملك المغرب؟	Who is the king of Morocco?
D 0173 AR AR	من هي زوجة محمد رضا البهلوي؟	Who is the wife of Mohammad Reza Pahlavi?
D 0174 AR AR	من أحرز على جائزة نوبل للسلام عام 2002؟	Who won the Nobel Peace Prize in 2002?
D 0175 AR AR	من أحرز على جائزة نوبل للسلام عام 1979؟	Who won the Nobel Peace Prize in 1979?
D 0176 AR AR	من بنى هرم خوفو؟	Who built the pyramid of Khufu?
D 0177 AR AR	من أفرج عن محمد نجيب؟	Who released Mohammed Najib?
D 0178 AR AR	ما هو اسم أخ سقمان؟	What is the name of the brother of Suqman?
D 0179 AR AR	من ألف"خواطر مصرحة"؟	Who is the author of "خواطر مصرحة"?
D 0180 AR AR	من هو رئيس سوريا؟	Who is President of Syria?
D 0181 AR AR	من اخرج فيلم"حرب العوالم"؟	Who produced "War of the Worlds"?
D 0182 AR AR	من كتب رواية"نهر سوميدا"؟	Who wrote the novel "Sumida River"?
D 0183 AR AR	من هي زوجة الرئيس الأمريكي جورج بوش الاب؟	Who is the wife of President George W. Bush?
D 0184 AR AR	من هو رئيس الوزراء السابق للأردن؟	Who is the former Prime Minister of Jordan?
D 0185 AR AR	من هو ملك الاردن؟	Who is King of Jordan?

D 0186 AR AR	من أسس مجلس قيادة الثورة في ليبيا؟	Who founded the Revolutionary Command Council in Libya?
D 0187 AR AR	من ألف كتاب"قصتي مع الشعر"؟	A book "My story with poetry"?
D 0188 AR AR	من حاز على جائزة نوبل الآداب عام 1988؟	Who won the Nobel Prize in Literature in 1988?
D 0189 AR AR	من هي زوجة انور السادات؟	Who is the wife of Anwar Sadat?
D 0190 AR AR	من هو الامام الرابع للشيعة؟	Who is the fourth Imam of the Shiites?
D 0191 AR AR	من هي زوجة علي بن ابي طالب؟	Who is the wife of Ali ibn Abi Talib?
D 0192 AR AR	بمن شبه الفرنسيون "فاطمة نسومر"؟	Who is the French equivalent of Fatima Nsumer?
D 0193 AR AR	من الذي تولى قيادة قوات المانيا وايطاليا في شمال افريقيا عام 1941؟	Who led the forces of Germany and Italy in North Africa in 1941?

Table 22: Table of Dataset Questions.

```
def similar_word(model,sentence,word,named_entities,central_words):
# find and print the most similar terms to a word
    val_sim_central_words=0.7
    val_sim=0.7
    if RepresentsInt(word) :
        an= number.ArNumbers()
    elif stopwords_classified.this_is_stop_word(word):
        return '' #Remove Stop words
    elif word in named_entities:
        return word #Option6--> fixé les entités nommées
    elif word in central_words:
        val_sim=val_sim_central_words;
    try:
        word_compose=word
        origin_word=word
        print origin_word
        word_compose=word_compose.split()
        if len(word_compose)>1:
            return word
        word=clean_str(word)
        most_similar = model.wv.most_similar( araby.normalize_ligature(word))
        lst=[x for x,y in most_similar if y>val_sim]
        lst=[''.join(c for c in x if c not in string.punctuation) for x in lst]
        if len(lst)==0:
            return origin_word
        lst=[seg.stem(x) for x in lst] #option10
        lst.insert(0,word) #insertion de mots source est obligatoire
        lst=set_list(lst)
        if len(lst)==1:
            return origin_word
        return '(' +(u' | '.join(lst))+')' #Option8--> similarwords to others u
    except Exception as e:
        return origin_word
```

Figure 53: Similar Words Function.