4-2020

Cross Language Information Transfer Between Modern Standard Arabic and Its Dialects – a Framework for Automatic Speech Recognition System Language Model

Tiba Zaki Abdulhameed
Western Michigan University, tlba.zakl@gmall.com

CROSS LANGUAGE INFORMATION TRANSFER BETWEEN MODERN STANDARD ARABIC AND ITS DIALECTS – A FRAMEWORK FOR AUTOMATIC SPEECH RECOGNITION SYSTEM LANGUAGE MODEL

by

Tiba Zaki Abdulhameed

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Computer Science
Western Michigan University
April 2020

Doctoral Committee:

Ikhlas Abdel-Qader, Ph.D., Chair
Imed Zitouni, Ph.D.
John Kapenga, Ph.D.
Robert Trenary, Ph.D.

CROSS LANGUAGE INFORMATION TRANSFER BETWEEN MODERN STANDARD ARABIC AND ITS DIALECTS – A FRAMEWORK FOR AUTOMATIC SPEECH RECOGNITION SYSTEM LANGUAGE MODEL

Tiba Zaki Abdulhameed, Ph.D.

Western Michigan University, 2020

Significant advances have been made with Modern Standard Arabic (MSA) Automatic Speech Recognition (ASR) applications. Yet, dialectal conversation ASR is still trailing behind due to limited language resources. As is the case in most cultures, the formal Modern Standard Arabic language is not used in daily life. Instead, varieties of regional dialects are spoken, which creates a dire need to address dialect ASR systems. Processing MSA language naturally poses considerable challenges that are passed on to the processing of its derived dialects. In dialects, many words have gradually morphed from MSA pronunciations and at many times have different usages. Also, a significant number of new vocabulary words have been imported from other foreign languages. In addition to these issues, dialects have low resources to be considered for any meaningful natural language processing (NLP) research. Therefore, there is a pressing need for an efficient language model (LM) for deployment in Arabic conversational speech recognition systems.

In this thesis, we explore building an Iraqi dialect conversational speech language model based on utilizing MSA data. Because there isn't a pre-defined annotated vocabulary set, our main approach is making use of word embedding for unsupervised clustering of the MSA-Iraqi

dialect words. Clustering the dialect words within the relative MSA words is employed to create a class-based LM. This allows the use of MSA data to cover the insufficiency of the dialect data. The model uses the dialect word's statistical history in addition to the statistics of related MSA words to make predictions of the intended spoken word sequence. Thus, efficient word embedding becomes important to produce a reliable LM.

To achieve efficient word embedding, first an analysis of the MSA and the Iraqi dialect vocabulary sets and their context intersection is conducted. For this purpose, Dialect Fast Stemming Algorithm (DFSA) is proposed that utilizes the MSA data and a predefined dialect suffixes set. The intersection set enlarged from 42.8% to 54% of the Iraqi vocabulary, and from 8% to 13% of the MSA vocabulary. Second, the syntax and semantic feature vector that is produced by applying the distributional-theory-based word embedding word2vec contained noise from having contexts that appear in MSA or in the dialect solely; thus, applying PCA reduced the perplexity (pp) by 6.7%. Finally, the novel Wasf-Vec topological word embedding algorithm is proposed, which relies on the hypothesis that for a rich morphological language like Arabic, the word's topological feature is of much significance to be considered. This new feature extraction technique addresses the high morphological properties and reduces PP by 7% when using distributional-theory-based word embedding. Moreover, a deep analysis of the words syntagmatic and paradigmatic relations are illustrated based on solid Arabic and Greek linguistic theories that prove the need of topological word embedding.

The three researches compiling this dissertation demonstrate the feasibility of utilizing MSA resources to enhance dialect processing. Further, combining distributional-theory-based and Topology-based word embedding is highly of great intense for future investigation.

ACKNOWLEDGMENTS

Gratitude for Allah, the most merciful and compassionate.

I would like to begin by expressing my great appreciation to the Higher Committee of Education Development (HCED) in Iraq for sponsoring this dissertation. Special appreciation goes to Dr. Abdulhakeem Alrawi, who sadly passed before I completed this work.

Also, I would like to show my gratitude to the LDC for the opportunity to use Arabic GALE phase2 part1 and part2 data through a data scholarship.

I am particularly grateful to my advisors Dr. Ikhlas Abdel-Qader and Dr. Imed Zitouni for their constant support and valuable advice. They wisely led and taught me how to think as a researcher and how to question and criticize research. In addition, they showed me what it means to possess a critical research mindset. Besides my advisors, I would like to thank the committee members Dr. John Kapenga and Dr. Robert Trenary for their expert input and insightful comments.

In addition, I wish to acknowledge the help provided by Mr. Robert Dlouhy, linguistics faculty member in the Department of World Languages and Literature at Western Michigan University, for his valuable discussions on linguistics terminology, and for Ms. Gari Voss, an adjunct professor in the English Department, for editing this dissertation. A special thank you is extended to the staff of the Computer Science Department at Western Michigan University, and my gratitude goes to Dr. Steve Carr, the Computer Science Chair, for the support he provided.

For ongoing support on this PhD journey, I would like to express my deepest appreciation to my husband for his patience, understanding and sacrifice. At the same time, my thanks go

to my sons, Ahmed and Ibraheem, for always encouraging me. For their continuous support, special appreciation is given to my father-in-law and mother-in-law. Finally, thanks to my beloved sisters for their warm love, and prayers.

Beyond my family, I thank my neighbors Osamma Al-Sharqi and Lubna Al-Abood for providing the transcript data from the Abu-flaies movie series. I feel blessed to have so many wonderful, supportive friends in my life and would like to thank all of them by name for their support, not only in my professional and academic career but in life, but there are not enough pages.

Last but not least, there are my parents. Thanks to my angel Mom for everything good in my life. She taught me how to keep thinking positively when facing challenges. I dedicate this dissertation to the soul of my father who was the first to encourage me to navigate this long journey. He inspired me to study computer science and search our natural language and specifically the Iraqi dialect. I am honored to have had such fine guidance through my life.

Tiba Zaki Abdulhameed

TABLE OF CONTENTS

ACKNOWLEDGMENTS ii

LIST OF TABLES vi

LIST OF FIGURES vii

ABBREVIATIONS ix

CHAPTER 1. INTRODUCTION 1
1.1 Arabic ASR 1
1.2 Issues in Arabic Dialect Languages 2
1.3 Automatic Speech Recognition Structure 3
1.3.1 Lexicon 4
1.3.2 Acoustic Modeling 4
1.3.3 Language Modeling (LM) 5
1.3.4 Search or Encoder 6
1.4 ASR Evaluation Metrics 6
1.5 Literature Review 7
1.6 Data Used 9
1.7 Data Preprocessing 9
1.7.1 Stemming 9
1.8 Main Goal 10
1.9 Thesis Structure 11
References 12

CHAPTER 2. ASSESSING THE USABILITY OF MODERN STANDARD ARABIC DATA IN ENHANCING THE LANGUAGE MODEL OF LIMITED SIZE DIALECT CONVERSATIONS 14
2.1 Introduction 15
2.2 Related Work 17
2.3 Language Model 18
2.4 Methodology 19
2.4.1 Preprocessing Data 19
2.4.2 Running word2vec 19
2.4.3 Combining Iraqi and GALE Corpora 22
2.4.4 Interpolating With LMs 23
2.4.5 Calculating Perplexity Using SRILM Tool 23
2.5 Experimental Results and Discussion 25
2.6 Conclusion 27
References 28

CHAPTER 3. ENHANCEMENT OF THE WORD2VEC CLASS-BASED LANGUAGE MODELING BY OPTIMIZING THE FEATURES VECTOR USING PCA 30
3.1 Introduction 31
3.2 Background 32

Table of Contents–Continued

3.2.1 Neural Word Embedding . . . 32
3.2.2 Principal Component Analysis (PCA) . . . 33
3.2.3 Class-Based Language Modeling . . . 34
3.3 Experimental Setup . . . 35
3.4 Results . . . 36
3.4.1 LM Perplexity (pp) . . . 38
3.4.2 Execution Time . . . 38
3.5 Conclusion . . . 40
References . . . 42

CHAPTER 4. WASF-VEC WORD EMBEDDING WASF-VEC: TOPOLOGY-BASED WORD EMBEDDING FOR MODERN STANDARD ARABIC AND IRAQI DIALECT ONTOLOGY . . . 43
4.1 Introduction . . . 44
4.1.1 Background and Problem Statement . . . 44
4.1.2 Word Relations . . . 47
4.2 Related Work . . . 49
4.3 Datasets Description . . . 53
4.4 Methodology . . . 53
4.4.1 Data Pre-processing . . . 54
4.4.2 Features Extraction . . . 56
4.4.3 Class-Based Language Modeling . . . 58
4.4.4 Analysis Method: Visualizing Feature Vectors . . . 59
4.5 Results and Analysis . . . 60
4.5.1 Dialect Fast Stemming Algorithm (DFSA) Accuracy . . . 60
4.5.2 Visualization Result of Feature Vectors . . . 61
4.5.3 Nearest Neighbors . . . 70
4.5.4 CBLM pp . . . 73
4.6 Conclusions . . . 73
References . . . 76

CHAPTER 5. CONCLUSION . . . 80
5.1 Contribution . . . 82
5.2 Future Work . . . 83

APPENDICES . . . 84
A Different Forms of the Word Give . . . 84
B Permission Letters . . . 85

LIST OF TABLES

2.1 Examples from classes.txt of words semantic clustering output of word2vec using CBOW gram of 10 cluster. 21

2.2 A comparison of the results of different LM mixing for multiple clustering methods. 24

3.1 Perplexity reduction ratio of low and high noise data. 36

3.2 Perplexity and clustering time using different vector size for low and high noise data.37

4.1 Examples of how paradigmatic and syntagmatic relations between words appear in MSA and the Iraqi dialect. 62

4.2 Sample of Wasf-Vec clustered words. 72

4.3 Sample of word2vec clustered words. 72

4.4 The pp of the CBLMs. 73

4.5 Comparing the Iraqi to MSA in terms of topological features. 74

LIST OF FIGURES

1.1 ASR Structure. W=w_1, w_2,..., w_n, where n is the context length. And O=o_1, o_2,..., o_t, where t is the number of frames in utterance. . . . 3

2.1 Steps diagram to produce statistical LM. . . . 20

2.2 Clustering and LM Generation. . . . 25

2.3 Samples of the Iraqi and GALE data. . . . 27

3.1 Counting and LM Generation [13]. . . . 34

3.2 Experimental setup flow diagram, either Features Vectors (FV) input directly to the class-based LM or PCA is applied to reduce the FV dimensionality, then input to class-based LM. . . . 36

3.3 Low noise data LM performance using different words feature vector length. . . . 39

3.4 High noise data LM performance using different words feature vector length. . . . 39

3.5 Clustering run time for different words feature vector length. . . . 40

3.6 Information in first third PCA components. . . . 41

4.1 Pronunciation similarity and semantic relations. . . . 52

4.2 Flow diagram of the steps for analyzing and evaluating words' features representations. . . . 54

4.3 Dialect fast stemming algorithm. . . . 55

4.4 Wasf-Vec: The words' topology features extraction. . . . 57

4.5 Illustration of Wasf. . . . 57

4.6 Iraqi-MSA Vocabulary Intersections. . . . 60

4.7 Plots of selected set1 from different regions of word2vec cloud is shown. (a) in word2vec (b) in Wasf-Vec. . . . 62

4.8 Plots of selected set2 from different regions of the Wasf-Vec cloud. (a) in word2vec (b) in Wasf-Vec. . . . 63

4.9 The lines represent the distances between paradigmatic related words in Wasf-Vec. 64

4.10 The lines represent the distances between paradigmatic related words in word2vec. . 65

4.11 The distances histogram of the paradigmatic related words. (a) Is word2vec space (b) Is Wasf-Vec space. . . . 67

4.12 The lines represent the distances between syntagmatic related words. (a) is word2vec space and (b) is Wasf-Vec space. . . . 68

4.13 The distances histogram of the syntagmatic relation of the words. (a) is word2vec space and (b) is Wasf-Vec space. 68

4.14 This figure shows how different '>nTa' rooted words are spread in the vector space. (a) is word2vec space and (b) is Wasf-Vec space. 69

4.15 This figure shows how words with the patterns fEl and fAEl are presented in the vector space. (a) is word2vec space, (b) is Wasf-Vec space. 69

4.16 This figure shows distance histogram of words with the patterns fEl and fAEl (a) is word2vec space, (b) is Wasf-Vec space. 70

4.17 This figure shows how words of pattern fEl and mfEwl are presented in the vector space. (a) is word2vec space, (b) is Wasf-Vec space. 71

4.18 This figure shows distance histogram of words with the patterns fEl and mfEwl (a) is word2vec space, (b) is Wasf-Vec space. 71

B.1 Permission to include "Tiba Zaki Abdulhameed, Imed Zitouni, Ikhlas Abdel-Qader, and Mohamed Abusharkh.Assessing the usability of modern standard Arabic data in enhancing the language model of limited size dialect conversations. Casablanca, Moroco, December 2017. International Conference on Natural Language, Signal and Speech Processing 2017." 85

B.2 Permission to include "Tiba Zaki Abdulhameed, Imed Zitouni, and Ikhlas Abdel-Qader. "Wasf-Vec: Topology-based Word Embedding for Modern Standard Arabic and Iraqi Dialect Ontology." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 19, no. 2 (2019): 1-27." 86

B.3 Permission to include "Tiba Zaki Abdulhameed, Imed Zitouni, and Ikhlas Abdel-Qader. Enhancement of the word2vec class-based language modeling by optimizing the features vector using pca. In 2018 IEEE International Conference on Electro/Information Technology (EIT), volume 2018-, pages 0866–0870. IEEE, 2018." 86

ABBREVIATIONS

MSA	Modern Standard Arabic
LM	Language Model
LDC	Linguistic Data Consortium
pp	Perplexity
POS	Part-of-speech-tag
NN	Neuron Network
HMM	Hidden Markov Model
SRILM	Stanford Research Institution Language Modeling
CMUCLMTK	Carengie Millon University Language Model Tool Kit
WER	Word Error Rate
H	Cross Entropy

CHAPTER 1. INTRODUCTION

Automatic Speech Recognition System (ASR) is the field of electrical engineering and computer science that is concerned with translating the audio signal to text. There has been increasing demand for efficient ASR in many applications such as computer aided learning, speech therapy (Assessment of Apraxia), meeting and conference summarization, information extraction and retrieval using voice search, speech to text translation, automatic indexing of audio files, facilitation of machine communication with people with special disabilities, and most recently gaming. All these applications, and others, have led to the development of Speech based User Interface (SUI). However, ASR systems are not perfect since humans have so many variations in their speech due to dialects, accents and unique ways of pronunciations of certain words. These issues increase the probability of having out of vocabulary words and incorrect prediction which require the speaker to repeat the word frequently until the machine is able to understand it. Thus, even with the existence of the current ASR, improving the accuracy would make the user more comfortable when using them.

1.1 Arabic ASR

Arabic is the sixth most spoken language and is spoken by more than a half billion people around the world. In addition to its religious importance for Muslims, it is one of the official languages of the United Nations and has gained increasing importance in political and economic fields. In recent years, the Arabic Language computational analysis domain has come to a promising stage [1]. According to the ProQuest database, statistics show that research in Arabic ASR has blossomed and grown since 2009.

In daily life communications, Modern Standard Arabic (MSA) is not spoken, but is the official language for education, media, and written documents. Thus, frequently in automated speech recognition applications, users prefer to communicate with the computer using simple dialectical forms of the Arabic language, for example, when giving orders to a robot or searching the web with dialectical spoken words. On the other hand, due to the lack of dialectical

language research and resources such as recorded and documented corpora, Arabic ASR systems are not easily usable if the user choses to speak in a dialectical form. For this reason, a mixed Language model of both MSA and dialectical Arabic is needed to make use of the MSR resources. Moreover, many real-life chatting sessions include English words. A one-hour lecture in computer science is a mixture of MSA, English, and dialectical languages. Thus, it is important to utilize the information identified regarding MSA and the English language ASR to develop a real life Arabic spoken ASR.

1.2 Issues in Arabic Dialect Languages

Arabic dialects have inherited rich morphology features from MSA. In addition to the morphology features inherited from the MSA, other issues related to the existence of new imported words and even new phonemes from non-Arabic neighbors during centuries of interaction, such as the Persian word *Aghatee* which in Iraqi dialect is used to show respect [2]. Nevertheless, people tend to choose easier, lighter word pronunciations in their conversations, so dialects contain many words that have incorporated the pronunciations of the original MSA word. The same MSA word can have a different usage in different dialects and this usage reflects cultural evolution as language is a symbolic communication of a culture [3].

In most of the Arabic dialects, the morphology has been changed a bit from the original MSA form. Therefore, some simpler morphological forms are used. For example, the feminine plural and dual verb forms have been merged with the masculine plural. Other morphological features have become more complex, such as having new prefixes such as *msh* مش and *mo* مو for negation, where the original interchanged MSA prefix is *mA* ما [4]. All of these issues have caused increases in the Out Of Vocabulary (OOV) words ratio and the sparsity issue in the data. For these reasons, a great deal of research needs to be done on the dialectical Arabic language.

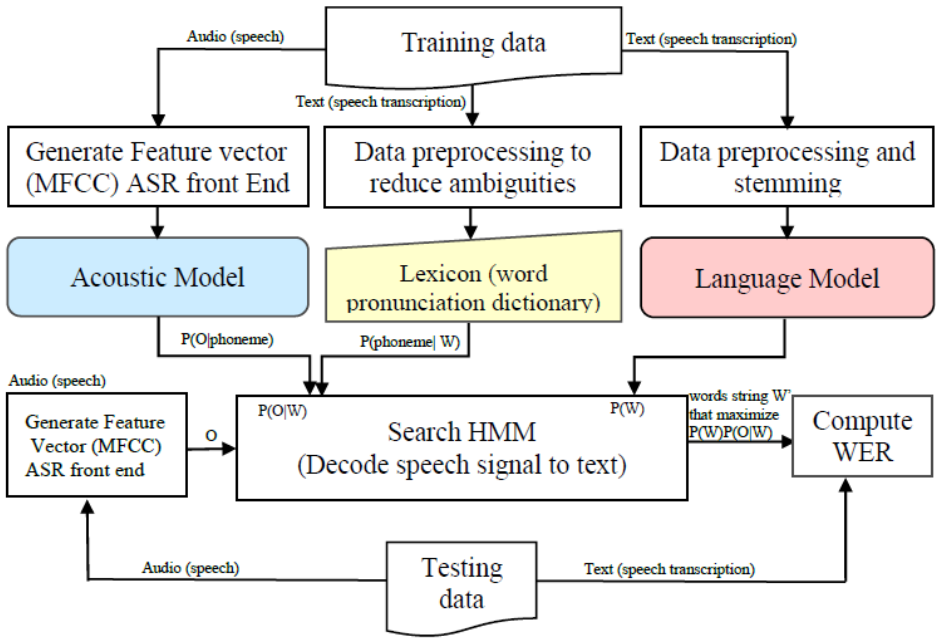


Figure 1.1.: ASR Structure. W=w_1, w_2,..., w_n, where n is the context length. And O=o_1, o_2,..., o_t, where t is the number of frames in utterance.

1.3 Automatic Speech Recognition Structure

The state-of-the-art structure of ASR is based on identifying phonemes as the smallest pronunciation unit. A speech utterance is a sequence of phonemes over a specific time period of speech. The utterance can be zero or more words. By zero, we mean an utterance like laughing, coughing, or even just noise. The end goal of ASR is to output an accurate transcript for the utterance.

Many components contribute to the accuracy of the ASR and work together. One inefficient component can affect the whole ASR performance. In our work, we are targeting the LM component because this component is language specific and is an important component in many other Natural Language Processing (NLP) applications.

The ASR structure is shown in figure 1.1 and its components are introduced in the following sections.

1.3.1 Lexicon

A lexicon is a file of words and pronunciations that identifies the sequence of phonemes or graphemes (sub-word units) for each vocabulary word. In other words, a lexicon builds a pronunciation model for each word. More than one sequence of phonemes for the same word can take place. In this case, the word may have different pronunciations. This is mainly important when considering conversational ASRs because the informal speech is free from restricted rules.

The Arabic lexicon can be constructed by direct mapping to the Buckwalter format of Romanizing the Arabic letters [5]. For example, *ghaiem* غائم, that means *cloudy*, is written in the lexicon as gA<m. A good lexicon needs to include other pronunciation occurrences of the same word such as if the word appears with different diacritic markings, for example, *ghaiemon* غائمٌ or *ghaieman* غائماً as gA<mN and gA<mAF, respectively. These pronunciation differences are mainly due to the short vowels that are presented in the Arabic orthography as diacritics.

1.3.2 Acoustic Modeling

A sub-word level recognition is done in the Acoustic model. The input audio signal is mapped to the Acoustic model based on either a Neural Network (NN) structure or on a Hidden Markov Model (HMM). The Acoustic model can be easily represented as a graph where the nodes are either NN nodes or HMM states. The Deep Neural Network (DNN) recently achieved good advances in this field of research [6]. The Acoustic model uses the lexicon graph to find the corresponding word to phoneme sequence. The decoder of the ASR uses the sequence scores to decide what path to follow where each path represents a given utterance.

1.3.3 Language Modeling (LM)

The next step after the Acoustic model nominates some words, the LM will predict the most coherent utterance spoken. One way to classify LMs is by considering them as either rule based or statistical. Rule based models are not as popular for complex languages due to the difficulties of reducing the ambiguities. Thus, the statistical Models depending on the probability of word combinations are the most commonly used LMs. The probabilities are estimated from the training data. In English, for example, a LM needs to give the most statistically appropriate word among the similarly pronounced words, such as there, their, or they're. An Arabic example is if the utterance phoneme was identified by the Acoustic model to be /gha<mmSHobzxAtmtr/. In this case, the LM will predict the most likely phrase or sentence among several.

- Cloudy sucking doom in falling coming rain غائم مص حوب بزخ ات مطر
- Cloudy with rain coming in falling غائم مصحوب بزخ ات مطر
- Cloudy with rain showers غائم مصحوب بزخات مطر

The last line is the most likely and coherent phrase.

Another example: if the utterance phoneme was identified by the Acoustic model to be /sharbimA>/ which will lead the LM to predict the most likely sentence among several.

- Devil in water شرَ بماء.
- He drank water شرَب ماء.

The last line is the most likely and coherent sentence.

From these examples, we can conclude that as the vocabulary set gets larger, more sentences can be formed from the same sequence of phonemes, and it will be more difficult for the LM to make the correct choice among the identified possibilities.

The most popular LM is the n-gram where the probability of a word is computed as the dependent probability of n-1 preceding words, as in Eq. 1.1.

$$p(utterance) = p(w_1)p(w_2|w_1)p(w_3|w_1w_2)...p(w_n|w_1...w_{\text{n-1}}) \tag{1.1}$$

However, in reality, it is time consuming to consider n-gram of n more than three or five and may not increase the efficiency of the LM. In general, a statistical LM can be either NN based or count based. Some LMs consider phonemes as the processing unit instead of whole words. Many alternatives LMs are clearly illustrated in [4] ch5. The choice of LM type depends on the language specifications.

The LM efficiency is measured in perplexity (pp). The pp is a way to quantify the cross entropy, which can be defined as uncertainty in a probability distribution [4] ch5. See Eq.1.2 Eq.1.3

$$H(P_{LM}) = -\frac{1}{n}\sum_{i=1}^{n} logP_{LM}(w_i|w_1...w_{\text{i-1}}) \tag{1.2}$$

$$pp = 2^{H(P_{LM})} \tag{1.3}$$

A smaller pp means greater LM efficiency. There are many ready to use tools to build a LM such as CMUCLMtk [7] and SRILM [8]. Both build a LM in ARPA format. There are other tools, but these two are the most standard and popular ones. The current research used SRILM in building and evaluating the various LMs.

1.3.4 Search or Encoder

The engine of the ASR is the decoder, which searches the best path in the HHM graph for any given input. The dominant algorithms used are the Viterbi decoder and Baum-Welch. The search space is mainly presented as the weighted finite state transducer or lattice. There are open-source ASRs that give the researcher the ability to use and make enhancements. The most popular ASRs are the Kaldi for C++ and other languages that can be integrated through bash scripting, the CMU-Sphinx for java programmers, and the HTK toolkit.

1.4 ASR Evaluation Metrics

A common way to measure the ASR's accuracy in transcription is to compute the Word Error Rate (WER). See Eq. 1.4 .

$$WER = \frac{S + D + I}{N} \tag{1.4}$$

where,

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- N is the number of words in the reference (N=S+D+C)
- C is the number of correct words,

1.5 Literature Review

One of the most important projects done on Iraqi dialect speech recognition and speech to speech translation was the TransTac program funded by the Defense Advanced Research Projects Agency (DARPA). The data and lexicon used in this project are supported by Linguistic Data Consortium (LDC) for official use only. Research on Iraqi speech and machine translation were done under this project. The data was enlarged in size through the time period of the project. Summarized advances made until 2009 were ended up with an ASR WER of 32% for a collected data set of l507 hours of Iraqi Arabic speech and text data using SRI's Dynaspeak® speech recognizer [9]. Morphological traditional class-based LM was proved to get the lowest pp. Other research focused on improving the acoustic model by considering the different word pronunciation in the lexicon ending with 2.4% reduction of WER [10]. This was done using the Janus Recognition Toolkit and 450-hour training set of a 62k word vocabulary set. Their research concluded with future recommendations to investigate pronunciation modeling in combination with discriminative training of acoustic models, and to investigate methods to optimally handle pronunciation variants of the same word within the language model and homograph issues.

For other Arabic dialects, Algeria's dialect was chosen by [11] and a new ASR named the Arabic Loria ASR system was introduced. The system used DNN with classical n-gram LM, and was tested for MSA with WER of 14%, while WER of 89% was recorded for 70 minutes of Algerian data. Adapting the acoustic model with both Arabic and French resources resulted in a WER reduction of 24% ending with a 65.45% WER. In addition, [12] investigated the use of MSA resources for the purpose of dialect ASR. A dialect independent phonemic acoustic model was introduced by first normalizing the MSA and dialect phonemes, then adapting the MSA phonemic acoustic model to the Egyptian dialect. A careful selection of 33 hours of MSA using the 34 phoneme-set and 0.5 hour of the 41 phoneme set was considered. An average reduction rate in WER was 18.2% using the CMU-Sphinx engine. For a more efficient pronunciation lexicon, automatically generated spelling variances for dialectal Arabic was used to solve the lack of transcribed dialect speech. The process was applied by adding the modified dialectal sound of the original Arabic orthographic letter such that one word can be defined in more than one pronunciation. For example, معقول have maEquwl and ma>uwl. In the context of improving the Acoustic model of MSA by automatically generating pronunciation of words, [13] produced phonemes based on their probabilities by keeping the most likely word pronunciation variants in the pronunciation's dictionary, which achieved a reduction of WER by about 1%. A different approach implemented by [14] depended on linguistic rules, MADA morphological analysis, and disambiguation tools, and achieved a 3.7%-7.29% reduction of WER.

In Kaldi ready recipe examples, there is the Arabic GALE [15] which uses MSA 203 speech hours and achieved 26.95% WER using the Triphone+DNN+MPE pipeline. Separating the conversational data and testing it alone achieved 32. 21% WER. The research showed the preference of using the MADA vowelization based phoneme system over the grapheme-based system. MADA vowelization followed by vowelization to phonetization (V2P) was the best approach used to generate phoneme lexicons with 35 speech phoneme and one silent phoneme, in addition to the Egyptian dialect callhome data set with best WER 52.29% using Triphone+SGMM+SAT+fMLLR pipeline. A Kaldi specific implementation results on a data set that is approximately similar in size or structure of the data set we are using is [16], who

developed the 18-hour Czech dataset for dialog-based ASR and achieved 48% WER using the Triphone+LDA+MLLT+ BMMI pipeline.

1.6 Data Used

The corpus used in our experiments is the Iraqi Arabic Conversational Telephone Speech (LDC2006S45) [17]. This corpus is taken from the Linguistic Data Consortium (LDC) and contains 276 Iraqi Arabic speakers in the form of Iraqi dialect telephone conversations. The dataset is subdivided into train-c1, train-c2, and devtest. Train-c1 represents one side of a recorded phone conversation and train-c2 is the other side. Both training sets are combined. The transcriptions consist of 199k words and its size is 1.8MB. The devtest is balanced in term of speaker diversity and account for 6% of the dataset. It is a certified standard test set according to the test process applied by the National Institute of Standards and Technology (NIST). Another 4% of the data was used as a tuning set. For our experiment, 90% of the corpus was used for training, that is, 199k words. The devtest was used for testing and included 102KB of about 12k words. In addition, the GALE dataset containing about 1516k words of MSA broadcast news and reports [18] [19] was also used for training.

1.7 Data Preprocessing

Data preprocessing needs to take place to reduce the sparsity of the vocabulary set. For Arabic data, we reduce the effect of the high morphology by applying MADAMIRA stemming tool and Iraqi stemming. In addition, other languages written words in the transcripts were removed. The third action was normalizing how Hamza is written, and removing stop words. Detailed explanations can be found in the following sections.

1.7.1 Stemming

Stemming was implemented using MADAMIRA-release-20170403-2.1 [20]. We applied additional Iraqi stemming through our proposed stemming algorithm, the Dialect Fast Stemming

Algorithm (DFSA) that does not need any additional tree-banks or Database. The DFSA does not need training because it depends solely on the vocabulary set and predefined suffix set. The objective of the proposed algorithm is to lower the data sparsity by reducing different word forms of similar stems.

For Iraqi stemming, the vocabulary set is a union of Iraqi and MSA. The vocabulary is extracted from the existing corpora. The algorithm mainly reduces the Iraqi specific prefixes from a word if the remainder of the word also exists in the vocabulary set. Words that will be under stemming consideration contain at least five letters, since words of less than 5 letters are rarely expected to be attached to prefixes. This is because most of the Arabic words' roots are 3 letters in length. The algorithm is fast because it does not consume learning time and can be applied on MSA by defining the MSA's expected suffix set. The algorithm is listed in Chapter 4 Algorithm in Figure 4.3 and the same procedure is applied for postfixes if needed. In Iraqi the postfixes did not need further processing since most of them were captured through MADAMIRA.

Applying both stemming techniques to the Iraqi and MSA data improved the ratio of the common words between their vocabularies. The stemming caused an increase in the intersection ratio between Iraqi and MSA words from 42.8% to 54.5% of the Iraqi vocabulary, as shown in Chapter 4. In addition, stemming reduced the vocabulary size of both Iraqi and MSA from 21k words to 13.4k words, and from 111k words to 53k words, respectively.

Stop Words

Words were eliminated using the nltk.corpus.stopwords.words('arabic') set in addition to a calculated set of the highest frequent words extracted from the Iraqi corpus.

1.8 Main Goal

The main goal of this dissertation is to propose an efficient Arabic dialect LM that can be used in building conversational ASR using MSA resources.

1.9 Thesis Structure

This dissertation systematically compiles the work findings documented in three published works and presented in Chapters 2, 3 and 4. Chapters 2 demonstrates how to build a word embedding class-based LM with analyses of the relation between MSA and Iraqi vocabulary and context to assess the ability of using MSA to enhance dialect LM. Chapter 3 addresses optimizing the feature vector length using PCA. Chapter 4 provides an in-depth analysis of MSA to Iraqi dialect words analogy captured by word embedding in addition to the new proposed features vector (Wasf-Vec). Finally, chapter 5 presents the conclusions and examines potential future work.

References

[1] Douglas R. Magrath. Bassiouney , reem, and graham katz e. (eds.). Arabic language and linguistics. Washington, DC: Georgetown university press , 2012 . pp. xiv, 232.44.95, paper.44.95, ebook. isbn 9781589018853. Modern Language Journal, 97(2):581–582,June 2013.

[2] Zaki Abdulhameed Alhabba, مُفرَدات فَارِسيَة في بَغداديات عزيز حجية [Persian vocabulary in Baghdadiat Aziz Hejiyah], 2002, Arab Encyclopedia House

[3] Ulrich J. Frey, Charlotte St¨ormer, and Kai P. Willf¨uhr, editors. Homo Novus, A Human without Illusions. The Frontiers Collection. Springer Berlin, Heidelberg, 2010.

[4] I. Zitouni. Natural language processing of Semitic languages. Theory and applications of natural language processing. Springer Berlin Heidelberg, 2014.

[5] Buckwalter. Buckwalter Arabic morphological analyzer version 1.0. 2002.

[6] Andrew L Maas, Peng Qi, Ziang Xie, Awni Y Hannun, Christopher T Lengerich, DanielJurafsky, and Andrew Y Ng. Building DNN acoustic models for large vocabulary speech recognition. Computer Speech Language, 41(C):195–213, 2017.

[7] Roni Rosenfeld and Philip Clarkson. Statistical language modeling using the cmu-cambridge toolkit. 1997.

[8] Andreas Stolcke. Srilm-an extensible language modeling toolkit. In Seventh international conference on spoken language processing, 2002.

[9] Murat Akbacak, Horacio Franco, Michael Frandsen, Sasa Hasan, Huda Jameel, AndreasKathol, Shahram Khadivi, Xin Lei, Arindam Mandal, Saab Mansour, et al. Recent advances in SRI's IraqComm^{TM} Iraqi Arabic-English speech-to-speech translation system. In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Con-ference on, pages 4809–4812. IEEE, 2009.

[10] Hassan Al-Haj, Roger Hsiao, Ian Lane, Alan W Black, and Alex Waibel. Pronunciation modeling for dialectal Arabic speech recognition. In Automatic Speech Recognition Understanding, 2009. ASRU 2009. IEEE Workshop on, pages 525–528. IEEE, 2009.

[11] Mohamed Amine Menacer, Odile Mella, Dominique Fohr, Denis Jouvet, David Langlois,and Kamel Smaili. An enhanced automatic speech recognition system for Arabic. In Proceedings of the Third Arabic Natural Language Processing Workshop, pages 157–165,2017.

[12] Mohamed Elmahdy, Rainer Gruhn, and Wolfgang Minker. Novel techniques for dialectal Arabic speech recognition. Springer Science Business Media, 2012.

[13] Frank Diehl, Mark JF Gales, Marcus Tomalin, and Philip C Woodland. Phonetic pronunciations for Arabic speech-to-text systems. In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, pages 1573–1576. IEEE, 2008.

[14] Fadi Biadsy, Julia Hirschberg, and Nizar Habash. Spoken arabic dialect identification using phonotactic modeling. In Proceedings of the eacl 2009 workshop on computational approaches to semitic languages, pages 53–61. Association for Computational Linguistics, 2009.

[15] Ahmed Ali, Yifan Zhang, Patrick Cardinal, Najim Dahak, Stephan Vogel, and James Glass. A complete kaldi recipe for building Arabic speech recognition systems. In Spoken Language Technology Workshop (SLT), 2014 IEEE, pages 525–529. IEEE, 2014.

[16] Matěj Korvas, Ondřej Plátek, Ondřej Dušek, Lukáš Žilka, and Filip Jurčíček. Free Eenglih and czech telephone speech corpus shared under the cc-by-sa 3.0 license. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), 2014.

[17] Sydney Appen, Pty Ltd and Australia. Iraqi Arabic conversational telephone speech LDC2006S45. Web Download. Philadelphia: Linguistic Data Consortium, 2006.

[18] Meghan Glenn and et al. GALE phase 2 Arabic broadcast conversation transcripts part 2 LDC2013T17. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

[19] Meghan Glenn and et al. GALE phase 2 Arabic broadcast conversation transcripts part 1 LDC2013T04. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

[20] Arfath Pasha, Mohamed Al-Badrashiny, Mona T. Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. Madamira: A fast,comprehensive tool for morphological analysis and disambiguation of Arabic. In LREC,2014.

CHAPTER 2. ASSESSING THE USABILITY OF MODERN STANDARD ARABIC DATA IN ENHANCING THE LANGUAGE MODEL OF LIMITED SIZE DIALECT CONVERSATIONS

"Tiba Zaki Abdulhameed, Imed Zitouni, Ikhlas Abdel-Qader, and Mohamed Abusharkh. Assessing the usability of modern standard Arabic data in enhancing the language model of limited size dialect conversations. Casablanca, Moroco, December 2017. International Conference on Natural Language, Signal and Speech Processing 2017."

Conversations are mostly spoken through a variety of dialects and the need for accurate speech recognition systems is growing in a wide range of applications. This paper presents an evaluation of the feasibility of using Modern Standard Arabic (MSA) data to enhance the perplexity (pp) of the Language Model (LM) for limited size Iraqi dialect conversations. We are interested in adapting the MSA's LM to Iraqi dialect by exploiting the capabilities of word embedding to deliver clusters of word vector representations classifying dialect specific words along with relevant MSA words. The vocabulary set of size 21k word extracted from Iraqi conversational speech from Appen's LDC was reduced to 14k word using the MSA-stemming followed by a dialect stemming technique. The stemming aimed to increase the intersection ratio between the dialect and MSA vocabulary, which are from GALE phase2 part1 and 2 from LDC. Various combinations of both corpora were tested. Two approaches were evaluated: 1) Separate word-based and class-based LMs, and 2) producing interpolated LMs of the various training data LMs. Results show that the second approach enhanced the only Iraqi LM around 8%, while improving the adaptation of the combined Iraqi and MSA LM to 30% using the interpolating technique. The results were promising when interpolating word based LMs with generalized LMs using word embedding feature vector classes based LMs. In addition, a 9% improvement of interpolating the class-based LM over interpolating word based alone LMs was achieved.

2.1 Introduction

Speech recognition tools have made major leaps in recent years and the accuracy of such technology grew from around 70% in 2010 to around 95% recently for market leaders like Google Now and Siri [1]. This can be attributed to a number of reasons but certainly robust LMs are critical to predict unclear sentences or sentence parts. However, producing efficient LM with low perplexity remains a challenge that needs addressing. This is especially true in the case of the Arabic language in which a complex lexicon and multiple living dialects make LM-based prediction a more challenging task.

A closer look at LM for Arabic dialects highlights the challenges that cause higher ambiguity than the standard language commonly termed MSA. Dialect pronunciation and rules differ from those of MSA. This can occur to the extent that these dialects can become mostly unrecognizable for MSA speakers so significant effort is required to facilitate communication between speakers of different dialects. Looking at the Iraqi dialect as an example, the Iraqi dialect has diversity in pronunciation based on a region spanning from the north to the south of Iraq, and many words originated from other regional languages such as Turkish, Farsi, and English. Also, some words come from MSA but have a slightly changed pronunciation. This causes an unbounded vocabulary set to develop. Moreover, speakers may use unrestricted grammar [2]. An additional challenge comes from the tendency of dialects to adjust the vocabulary and language usage during relatively short time intervals to reflect the cultural and generational changes. As an illustration, a new expression for "a friend" was introduced by one popular Iraqi T.V. comedy series. This led many young people to use the new term, and thus, a new vocabulary synonym of close friend was introduced to the daily conversations. Such word usage varies over time and LM should be enabled to recognize such evolution and adapt.

Within our assessment of the usability of MSA in enhancing the Iraqi dialect LM, we addressed three main challenges for the Iraqi dialect conversations LM adaptation; data sparsity, adapting different speech domains (conversations vs. broadcast news), and the data size limitation. Resolving these issues would result in a more accurate LM.

The first issue of the data sparsity, which is a feature inherited from the mother language MSA is our first challenge and the most difficult one. We resolved the data sparsity issue by using a class-gram LM and employing word2vec [3] in a comprehensive scheme that produces class n-gram. The word2vec program is introduced as a machine learning tool that could construct feature vectors for words in the input data set [3]. It has been successfully used in sentiment analysis and document classification work [4], [5]. Since we are using both MSA and Iraqi, words appears in many different contexts. We hypothesize that employing word clustering based on a k-means classifier can produce effective language modeling when words are used in various ways. Thus, clustering words and using class probability, where the cluster number is the word's class, would reduce the sparse words probability. More specifically, we propose using word2vec to cluster corpus words into classes that, in turn, would be used to build a language model using class n-gram technique [6]. The second challenge related to the fact that we are targeting conversational speech transcription language modeling and attempting to produce improvements by adapting MSA broadcast news and reports LM to Iraqi dialect phone conversations. We resolved this issue by interpolation of both data sets' LMs.

The third challenge is due to the small size Iraqi training data, which we resolved by expanding our training data with data taken from MSA corpus. This work is focused on exploring the feasibility of mixing the training data to enlarge the limited size dialect data.

We define this problem as an adaptation problem and consider Iraqi dialect as domain specific, while MSA is the general big background data set. Our scheme has the objective of minimizing LM perplexity by adapting the MSA LM to the Iraqi dialect, which can lead to significant improvements in the speech recognition systems for dialect-based conversations.

Chapter 2 is organized as follows. Section 2.2 discusses the most pertinent research efforts. Section 2.3 presents the language model used in this work while section 2.4 includes the experimental proposed scheme, and 2.5 relays the setup, results, and analysis. Finally, Section 2.6 is a conclusion of this chapter.

2.2 Related Work

A class-based LM was shown to be effective in solving the data sparsity problems of datasets. Instead of depending on independent word prediction, words are clustered into classes and via modeling, prediction can be achieved. This was shown in [7] where Part Of Speech (POS) techniques were used to classify the data and produce the neural network (NN) n-gram LM. Previous efforts considering Egyptian dialect LM for ASR reported that the best prediction results are achieved when an n-gram is combined with a class n-gram [8].

The challenge in this scenario rises from the assumption that the data set is fully classified. Yet, most available dialect data sets are not annotated for the purpose of word classification. A major data classification method that has proven to be effective for other languages is the count base statistical clustering of the corpus such as the hierarchical Brown clustering. NN-based word embedding was also used to replace manual tagging of POS. In [9], it was shown that word embedding can be used for unsupervised POS tagging.

In the context of the Arabic language, word2vec was used as a word embedding tool for Arabic sentiment analysis in [10]. They considered MSA and the dialectal Arabic sentiment opinion (specifically Egyptian dialect). The features were extracted from word2vec as an alternative to hand-crafted methods. In addition, word2vec proved its applicability in Arabic information retrieval and short answer grading in [11]. This was tested using twitter and book reviews that are considered a combination of more than one dialect, while new articles were considered as MSA. Also, word2vec was used to produce a comparable corpus (CALLYOU) from Youtube for Algerean, MSA, and French comments [12].

Linear interpolation has been proposed in many scenarios as an effective method of LMs adaptation. In the literature, the adaptation of domain specific LM of the same language was mainly explored [13] [14]. In a similar way, we are considering the dialect as a separate language domain that has intersection with MSA domain. For mixing heterogeneous text data of domain-specific LM, [15] produced term weighting that is used to decide in-domain and out-of-domain text segments for the purpose of document classification. This inspired us to produce a new weighting function to filter the MSA from text segments of less common words

with the dialect.

Our approach aims at making use of the semantic representation of the word embedding in word2vec for generating class-ngram. It is noted in the literature that most of the efforts are focused on MSA with a few efforts considering Egyptian and Levantine dialects. Despite being rich and commonly used, the Mesopotamian dialect group is spoken by approximately 30 million people, yet there is a lack of research catering to this prominent dialect. Thus, there is a need for a comprehensive LM that improves speech recognition performance and potentially highlights dependencies between the Iraqi dialect and MSA.

2.3 Language Model

One of the main factors in efficient ASR is the language model that will be fed to the system to decide the hypothesized spoken word using context-based probabilities. The pp is the metric that is used for computing the Language model efficiency [6]. A lower value of pp means high value of the expected word probability in a certain context and naturally a lower number of bits needed to encode the words. This enables easier decision making (i.e. reduced sparsity) [8].

The simplest possible regular LM is the unigram. Each unit of the language has a probability of its count divided by total words count in the corpus. An n-gram model is a model that counts the word probability by taking into consideration word history (i.e. context) of length n-1. So, the probability of a certain word x to appear in the corpus given the previous word was word y is

$$p(x|y) = \frac{count(yx)}{count(y)} \tag{2.1}$$

Certain algorithms can be applied to smooth the probabilities and achieve better pp such as Kneser-Ney [8]. Smoothing includes three operations to improve accuracy, namely, discounting, back-off and interpolation. For example, a class based bigram computes the probability of word x, given the previous word, y as

$$p(x \mid y) = p(x \mid class(x))p(class(x) \mid class(y)) \tag{2.2}$$

Where the conditional probability of the word x, given that it appears in a unique class, class(x), is defined as the ratio of the number of occurrences of word x, to the total number of occurrences of its class within the corpus. Classes are mutually exclusive where a word belongs to one class only. In addition, classes lengths are not necessarily equal.

$$count(class(y)) = \sum_{i \in V} count(w_i \in class(y)) \tag{2.3}$$

Where V is the vocabulary set

$$p(x \mid class(x)) = \frac{count(x)}{count(class(x))} \tag{2.4}$$

$$p(class(x) \mid class(y)) = \frac{count(class(y)class(x))}{count(class(y))} \tag{2.5}$$

2.4 Methodology

2.4.1 Preprocessing Data

Data preprocessing includes removal of unneeded tags followed by the preprocessing stemming steps described in Section 4.4.1. OOV words are words that are not in the Iraqi word vocabulary set.

2.4.2 Running word2vec

As shown in Fig. 2.1, the preprocessed training dataset is used as an input into the word2vec tool. word2vec uses a single hidden layer that is fully connected NN. The input and output layers are both of the same cardinality of the vocabulary. To reduce computation complexity, the hidden layer was replaced by simple projection. This makes word2vec capable of large data training. The feature vector is extracted from the weights matrix between the input layer and the hidden layer where the end target would be the word's neighbors in context [3]. A continuous Bag Of Words (CBOW) was tested to predict a word from the input

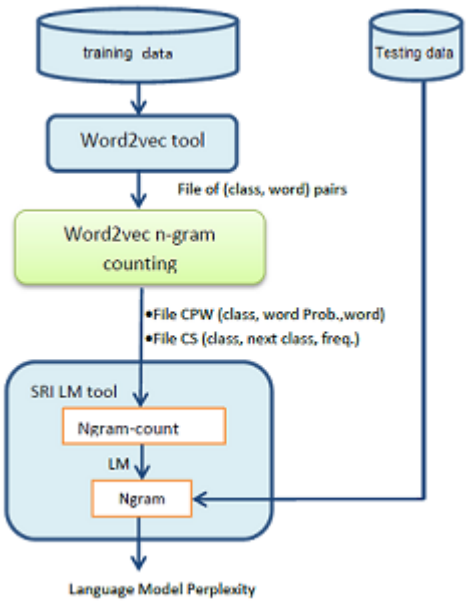


Figure 2.1.: Steps diagram to produce statistical LM.

context window by applying word2vec with 700 class, vector size 350, and window width of 5.

Multiple output produced by word2vec is then used in describing the language model. In our work, we are more interested in the output of the word2vec (classes.txt) file which contains class-word pairs for all of the words in the dataset. To show how word2vec semantically classifies the data, we present a few examples of word classes from the corpus in table 2.1. For example, class 3 shows some words of the same root clustered together, which clearly is reflecting the semantic and syntactic clustering dependency. In class 0, we also notice the semantically related words clustered together. In class 9, three different orthography formats of the same word اهلاً, which means 'Hi or welcome', appeared in the same class. This implies that the word2vec was to a certain extent successfully capturing semantic and syntactic relations of Iraqi dialect.

Next, we built a supporting tool in Python that takes one file (classes.txt) from word2vec output and generates two files, CPW and CS, with the Iraqi unclassified words separated in one class of its own. Also, each line in CPW contains the tuple class C, probability P of a word within class, and the word W. P was computed per equation 2.4.

In the CS file, each line contains the tuple: (class C_i, next class C_j, the number of occur-

Table 2.1.: Examples from classes.txt of words semantic clustering output of word2vec using CBOW gram of 10 cluster.

class9	class3	class6
السلام greeting peace	سوية together	مشغولة busy
اهلاً greeting Hi	تجي come	ملتهي busy
أهلاً greeting Hi	تروح go	مريضة sick
اهلا greeting Hi	ترجع return	مشاكل problems

rence of any word of class C_i followed by any word of class C_j for all classes i and j). The "start" of the sentence is represented using the tag <s> while the sentence end is indicated by the tag < /s >. We used two approaches to handle the out of vocabulary (OOV) words. In the first, we tagged those words as < unk > and they were treated as words in their own unified class while in the second they were just left as is and classified with the vocabulary words but were not considered in computing the probabilities for the language model.

CS is then used with SRILM to count class n-gram LM, while CPW is used in computing the pp as follows:

See the following command:

```
$ ngram-count -order 2 \
  -read w2vClassBasedLM-CS -write f1.ngrams

$ ngram-count   -order 2 \
 -read f1.ngrams -lm  w2vClassBasedLM.lm
```

To calculate the pp,

```
$ ngram -lm w2vClassBasedLM.lm \
 -classes CPW -ppl test-data
```

To calculate the interpolated LMs pp,

For example w2vClassBasedLM1.lm and w2vClassBasedLM2.lm,

```
# We first need to differentiate the classes in CPW files of the
  interpolated class based LMs by assigning some identification letter
  to the classes names. Here we used letter 'G'

$sed  -e 's/^/G/' w2vClassBasedLM2-CPW > w2vClassBasedLM2-CPW2

#Then concatenate both CPW files for each LM

$cat w2vClassBasedLM2-CPW2 w2vClassBasedLM1-CPW > combined_classes.CPW
#make the same changes of classes names on the LM
$sed -e 's/CLASS/GCLASS/g' \
-e 's/\<s\>/G\<s\>/g' \
-e 's/\<\/s\>/G\<\/s\>/g' \
w2vClassBasedLM2.lm  >  w2vClassBasedLM2.G.lm
#Interpolate them and assign the lambdas
L=0.5
L2=0.2
 $ngram  -unk  -lm LM1.lm -lambda $L\
  -mix-lm  w2vClassBasedLM1.lm  \
  -mix-lm2 w2vClassBasedLM2.G.lm  -mix-lambda2 $L2\
  -classes combined_classes.CPW -bayes 0 \
  -ppl $test -write-lm New.lm
```

2.4.3 Combining Iraqi and GALE Corpora

To compensate for the small data size that is available for this project, we combined Iraqi with the GALE corpora. This was performed via two methods as follows:

- Set1: Iraqi dialect phone calls corpus [16]. Testing data are the Devtest data defined as part of this corpus from the Linguistic Data Consortium (LDC).
- Set 2: Set 1 with 10% of Set 3.
- Set 3: GALE MSA [17] [18], which is broadcast news and report corpus.
- Set 4: 10 times duplicated Set 1 combined with Set 3. This will enlarge the frequency of the Iraqi context.
- Set 5: Iraqi and the whole of GALE, i.e Set 1 and Set 3.

In addition, refining the GALE data was also considered for which we computed, for each sentence, the probability of the OOV words that are tagged as <unk>. If the probability is more than 0.3, then the sentence will be discarded and considered noise.

2.4.4 Interpolating With LMs

The interpolation is presented using the following equation:

$$\begin{aligned} p(w) &= 0.5p_{IraqiTri-gram}(w) + 0.3p_{IraqiCB}(w) \\ &+0.2p_{GALECB}(word) \end{aligned} \tag{2.6}$$

where CB refers to word2vec class n-gram. The lambda weights were estimated using the Iraqi data tuning set. The lambda weights' combination that produced best results for the tuning set were chosen for the final interpolated versions. The flow diagram of the interpolation that produced the best results is shown in Fig.2.

2.4.5 Calculating Perplexity Using SRILM Tool

The SRI tool for Language Modeling [19] was introduced as a solution to generate LM and compute perplexity of the test data. All tests were unified on the same Iraqi devtest and the same vocabulary set.

Table 2.2.: A comparison of the results of different LM mixing for multiple clustering methods.

LM	pp	LM	pp
Set 1 tri-gram	124.2	Set 1 CB	136.5
Set 3tri-gram	636	Set 3 CB	509 467
Refined Set 3 tri-gram	628.8	Refined Set 3 CB	471.8
Set 5 tri-gram	177.6	Set 5 CB	367.8
Refined Set 5 tri-gram	176	Refined Set 5 CB	342.7
set 2 tri-gram	137.24	set 2 CB	185.2
Refined Set 2 tri-gram	136.764	Refined Set 2 CB	169.9
Set 4 tri-gram	208.3	Set 4 CB	282.4
Refined Set 4 tri-gram	207.7	Refined Set 4 CB	286.9
Interpolated Set 1 CB, and Set 1 tri-gram	114		
Interpolated, Set 2 , and Set 1 tri-gram	130.8	Interpolated, Set2 CB , Set 1 CB, and Set 1 tri-gram	116
Interpolated Set 3 tri-gram, and Set 1 tri-gram	148	Interpolated Set 1 CB, Set 3 CB and Set 1 tri-gram	119.5
Interpolated Set 5 tri-gram and Set 1 tri-gram	130	Interpolated Set 5 trigram, Set 1 CB, and Set 1 tri-gram Interpolated Set 5 CB, Set 1 CB, and Set 1 tri-gram	124 118

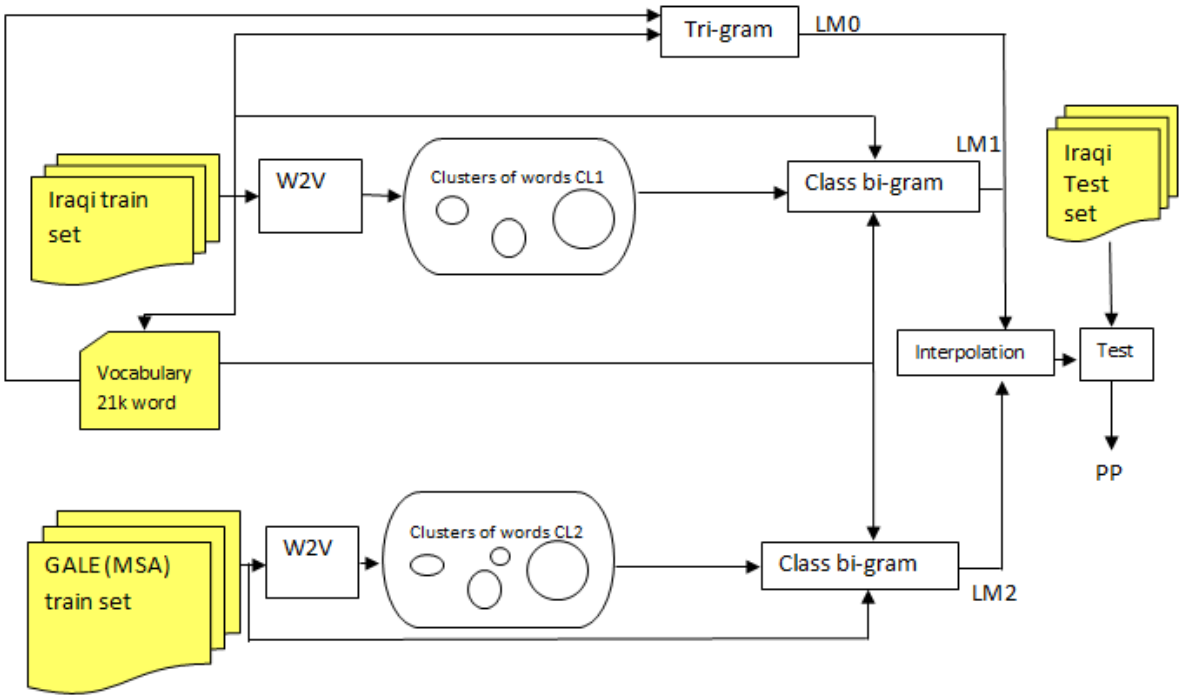


Figure 2.2.: Clustering and LM Generation.

2.5 Experimental Results and Discussion

The upper part of table2.2 shows results obtained from word trigram LM on the left side, and the word embedding class based LM on the right side, while in the lower part, we present interpolated versions of some of the upper LMs. In our attempt to expand the Iraqi data by mixing it with GALE, one might not expect that only about 54% of Iraqi words will actually appear in the MSA. This was clear when we extracted the intersecting words between them. Surprisingly, we found that the intersection set is only 7.3kword after stemming, yet only about 3.2kword is considered for clustering using word2vec. This is because in our experiment, we ignored words of frequency less than five from classification. This is similar to the default setup in word2vec to ensure good training of feature vectors. This may justify the lack of improvements under the method of only mixing two corpora. Indeed, Iraqi dialect and MSA are both Arabic, but they have almost different vocabularies which prevented this method from performing. Also, the data nature, that is news and broadcast, is different in context and vocabulary choices from those in phone conversations.

OOV words are either kept as they are or replaced by the <unk> to enable refining in a later step. The refining of the data, based on the <unk> ratio in the sentence, produced better results in the class-based LM, while it did not improve the word based trigram more than 1. This indicated the sensitivity of the word embedding class based method to the noisy sentences.

As an illustration, one arbitrary segment of each corpus is further explored in Fig. 2.3(a), 2.3(b), 2.3(c), and 2.3(d). The sample segments are for both Iraqi and GALE, and both are presented before stemming and after stemming. One can notice that <unk> words appeared in GALE due to the use of only Iraqi vocabulary. These <unk> words refer to word appearing in GALE but not in Iraqi. This causes high ambiguity when testing the Iraqi devtest data. In fact, the only common words in this arbitrary segments are أنف لا ما ها, where most of them are connecting words except أنف. These connecting words do have high frequency in both corpora, which means that the gain of having them in GALE was not that significant in enhancing the LM. We need to support words that appear in an average frequency in the Iraqi data.

Though, refining GALE data before calculating word embedding class based LM achieved on average 5% improvement over the LMs trained on unrefined MSA data, interpolating word-based with class-based of the Iraqi and any other set that contains MSA, produced relatively near results due to the λ weights that were 0.5, 0.3, and 0.2 for the Iraqi alone tri-gram, Iraqi CB, and (other MSA Set) CB LM respectively.

Interpolating the language models and enforcing the MSA's LM contributions to be small via the small λ weight leads to sharing common word probabilities and reducing the sparsity.

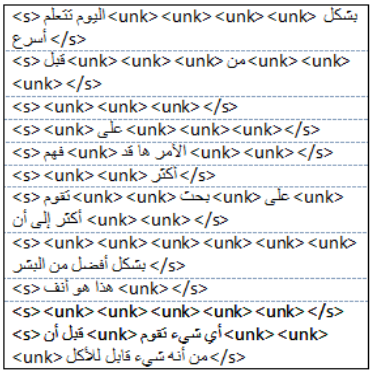
Also, we would like to include a note about the classification of words that are in Iraqi but not in GALE. Theoretically, the class probability would enhance the conditional (word|class) probability for these words; however, discarding all the GALE MSA vocabulary did not allow for this enhancement to occur. Nevertheless, if we consider all the GALE vocabulary, the sparsity of the words will be higher causing a higher pp.

We have actually considered the MSA words in the vocabulary, and the results show that not all Iraqi words clustered within similar MSA words, but there are some that did, such as

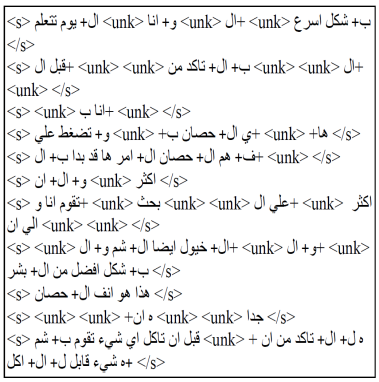
words with different spellings but the same actual use and syntax.

(a) Sample of Iraqi data.

(b) Sample of Stemmed Iraqi data.



(c) Sample of GALE data.



(d) Sample of Stemmed GALE data.

Figure 2.3.: Samples of the Iraqi and GALE data.

2.6 Conclusion

We attempted to investigate the value of using GALE MSA data to reduce the Iraqi dialectal conversation ambiguity through mixing the data and interpolating word-based with word-embedding feature vector classes-based language models. LMs interpolation achieved encouraging improvements and so future consideration can be given to such an approach. In addition, a different treatment for OOV words can be considered for future study. Also, a biased selection of a subset of GALE's dataset can be used to enhance the results further.

References

[1] K. Ryan. Who's smartest: Alexa, siri, and or google now. Inc. Magazine (2016).

[2] Sherri Condon, Mark Arehart, Dan Parvaz, Gregory Sanders, Christy Doran, and John Aberdeen. Evaluation of 2-way Iraqi Arabic–English speech translation systems using automated metrics.Machine translation, 26(1):159–176, 2012.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. 2013.

[4] Cícero Nogueira Dos Santos and Maira Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In COLING, pages 69–78, 2014.

[5] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In International Conference on Machine Learning, pages 957–966,2015.

[6] Frederick Jelinek. Statistical methods for speech recognition. MIT press, 1997.

[7] Ahmad Emami, Imed Zitouni, and Lidia Mangu. Rich morphology based n-gram language models for Arabic. In Ninth Annual Conference of the International Speech Communication Association, 2008.

[8] I. Zitouni.Natural Language Processing of Semitic Languages. Theory and Applications of Natural Language Processing. Springer Berlin Heidelberg, 2014.

[9] Chu-Cheng Lin, Waleed Ammar, Chris Dyer, and Lori S. Levin. Unsupervised POS induction with word embeddings. CoRR, abs/1503.06760, 2015.

[10] A. Aziz Altowayan and Lixin Tao. Word embeddings for arabic sentiment analysis. In2016 IEEE International Conference on Big Data (Big Data), pages 3820–3825. IEEE,2016.

[11] Mohamed A Zahran, Ahmed Magooda, Ashraf Y Mahgoub, Hazem Raafat, Mohsen Rashwan, and Amir Atyia. Word representations in vector space and their applications for arabic. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 430–443. Springer, Cham, 2015.

[12] Karima Abidi, Mohamed-Amine Menacer, and Kamel Smali. Calyou: A comparable spoken algerian corpus harvested from youtube. 08 2017.

[13] Xunying Liu, Mark John Francis Gales, and Philip C Woodland. Use of contexts in language model interpolation and adaptation. Computer Speech Language, 27(1):301–321,1 2013.

[14] Jerome R Bellegarda. Statistical language model adaptation: review and perspectives.Speech communication, 42(1):93–108, 2004.

[15] Ján Staš, Jozef Juhár, and Daniel Hládek. Classification of heterogeneous text data for robust domain-specific language modeling. EURASIP Journal on Audio, Speech, and Music Processing, 2014(1):14, Apr 2014.

[16] Sydney Appen, Pty Ltd and Australia. Iraqi Arabic conversational telephone speech LDC2006S45. Web Download. Philadelphia: Linguistic Data Consortium, 2006.

[17] Meghan Glenn and et al. GALE phase 2 Arabic broadcast conversation transcripts part 2 LDC2013T17. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

[18] Meghan Glenn and et al. GALE phase 2 Arabic broadcast conversation transcripts part 1 LDC2013T04. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

[19] Andreas Stolcke. Srilm an extensible language modeling toolkit. In Seventh international conference on spoken language processing, 2002.

CHAPTER 3. ENHANCEMENT OF THE WORD2VEC CLASS-BASED LANGUAGE MODELING BY OPTIMIZING THE FEATURES VECTOR USING PCA

" Tiba Zaki Abdulhameed, Imed Zitouni, and Ikhlas Abdel-Qader. Enhancement of the word2vec class-based language modeling by optimizing the features vector using PCA. In 2018 IEEE International Conference on Electro/Information Technology (EIT), volume 2018-, pages 0866–0870. IEEE, 2018. "

Neural word embedding, such as word2vec, produces very large feature vectors. In this Chapter, we are investigating the length of the feature vector aiming to optimize the word representation results, and also to speed up the algorithm by addressing noise impact. Principal Component Analysis (PCA) has a proven record in dimensionality reduction so we selected it to achieve our objectives. We also selected class based Language Modeling as extrinsic evaluation of the features' vectors and are using Perplexity (pp) as our metric. K-means clustering is used to classify words. The execution time of the classification is also computed. As a result, we concluded that for a given test data, if the training data is of one domain then large vector size can increase the precision of describing word relations. In contrast, if the training data is from different domains and contains large number of contexts not expected to occur in the test data then a small vector size will give a better description to help reducing the noise effect on clustering decisions.

Two different data training domains were used in this analysis; Modern Standard Arabic (MSA) broadcast news and reports, and Iraqi phone conversations with testing data of the same Iraqi data domain. Depending on this analysis, same domain training data and test data have execution times reduced by 61% while keeping same representation efficiency. In addition, for different domain training data i.e. MSA, pp reduction ratio of 6.7% is achieved with time reduced by 92%. This implies the importance of carefully choosing feature vector size on the overall performance.

3.1 Introduction

Vector representation of words can be produced using many methods, but the most common one is word2vec which is a neural network-based word embedding technique. To use word2vec, one needs to select certain parameters such as context window size, sub-sampling rate, and the length of the feature-vector. In the original paper, that produced word2vec [1], vector size of 300 was used. We concluded that these parameters are very crucial decisions and differ from one problem to another but did not reveal how to decide on the values of the most of the parameters. In this Chapter, we investigate how to optimize the length of the feature vector and the impact it has on performance. It is clear that the feature vector length is a very important parameter that not only impacts the accuracy of word representation, but also can influence runtime of word2vec and the execution time based on the applications.

Parallelization of word2vec training was applied by facilitating GPUs to speed up the run time [2]. The parallelization needed to define the dependency in its algorithm. Algorithm dependency is the main issue in parallelizing training algorithms like word2vec. Thus, optimizing these algorithm's word feature vector length aiming for better performance by analyzing words feature vectors is our focus in this Chapter.

Originally, word2vec applies negative sampling for noise defined as randomly generated contexts of words to increase the accuracy of semantic relations between words occurring in same context [1]. The injected noise is given very low weights in order to bring up the true contexts' weights. We define noise in the training data as contexts that are not expected to occur in the test data. It is also worth noting that eliminating the effect of Out of Vocabulary Words (OOV) to focus on context impact is done by replacing such words with Unknown Word Symbol (unk) in the training set. We also investigate the potential of using PCA of the words feature vectors to analyze the vector length on the performance of the LM and compute time. Other related works can be found in [1], were they focused on extracting the first 2 components of PCA to allow for words visualization in 2D space. On the same context, [3] showed that the first 2 components of PCA can also allow us to explore the most variant relations between words, which is the semantic meaning, while other feature reduction techniques can capture

less variant relations between same semantic group of words, such as the syntactic relations of verb tense. In [4] PCA was tested on 40 languages and demonstrated a better performance than a skip-gram model. Research results indicate that a skip-gram can be presented as an application of the exponential-family principal components analysis (EPCA). Also, [5] concluded that for efficient word representation, negative sampling is not enough, and we need to apply Noise Contrastive Estimation (NCE). Mathematically, [6] showed that word2vec Skip-Gram with negative sampling is actually a weighted logistic PCA. The word2vec was applied to a dialog act recognition task and it was reported that regardless of the size of the training corpus, word2vec could not capture valuable information [7]. Their reported limitations of word2vec was based on investigations of corpus size but not the word2vec model parameters. In our research, we find that the corpus size is not the only factor affecting the choice of the feature vector length but also the number of unexpected contexts that is embedded in the training set only. This is in agreement with [8], where a thorough comparative analysis of various word embeddings is implemented to address the problem of generating best word representations. Although, the study recorded similar behavior of various word embeddings in response to vector dimensionality, they could not reach a suggestion of the best choice for the vector length. They found that the domain of the corpus had more impact than the corpus size. Such research findings propelled our motivation to investigate the feature vector dimensionality.

3.2 Background

3.2.1 Neural Word Embedding

As the size of the training corpora increased with the availability of web resources, the need for relatively fast unsupervised algorithms to extract word relations rapidly increased. Research in Natural Language Processing (NLP), especially for language modeling tasks, improved greatly by making use of words feature vectors that are automatically generated using fully connected neural net techniques.

word2vec [9] is a tool that produces large dimension vectors for large vocabulary size. It

is a feed forward Neural Network where the connections weights between the input layer and the hidden layer construct the feature vector of each word. Both the input and output layer contain nodes equivalent to the number of words in the vocabulary set. A user can choose either a skip gram model or a continuous bag of word model. The type of the model decides what node would be highlighted in the input and output layer. Continuous bag of words tries to predict a word given its surrounding context, while Skip-gram predicts the context within a word may occur.

3.2.2 Principal Component Analysis (PCA)

While having a high dimension representation of words, a projection onto a lower dimension is needed to better interpret the feature vectors and to achieve improved performance. We also aimed to better our understanding and reveal the underlying structure of the data, which is very important for subsequent language modeling and word prediction.

Assume N is the number of words in the corpus, and each word feature vector is of length d, in order to reduce it to k, where k<d, features covariance (or correlation) matrix R with size (d×d) is constructed. Thus, the covariance σjl between two features j and l is computed using

$$\sigma jl = \frac{1}{n-1}\sum_{i=1}^{N}(x_{ij} - \mu_j)(x_{il} - \mu_l) \tag{3.1}$$

where, x_{ij} is the j^{th} feature of word i. To summarize, to produce the matrix R, apply (3.2)

$$R = \frac{1}{n-1}(X - \vec{\mu})^T(X - \vec{\mu}) \tag{3.2}$$

Then eigenvalues and eigenvector are constructed from the covariance matrix R. The end goal is to project the original data feature matrix, that is of (N×d), using the best k eigenvectors as its axes. To choose a number k, the first k eigenvectors should maintain 99% of the cumulative variance. This is done by choosing k eigenvectors corresponding to the k highest eigenvalues in a projection matrix P of size (d×k). Then recreate original data but with

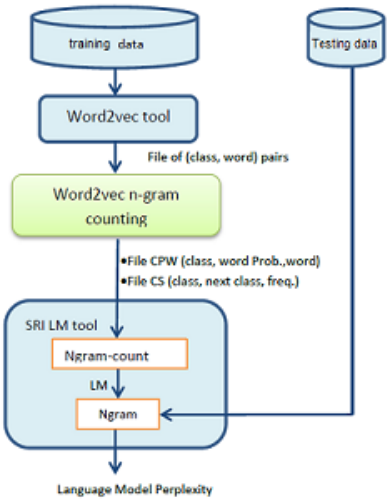


Figure 3.1.: Counting and LM Generation [13].

reduced dimensionality by multiplying original data feature matrix by the projection matrix, which will be of size (N×k). This reduction preserved the most important features that hold most of the information [10].

3.2.3 Class-Based Language Modeling

Evaluating the word2vec would be either extrinsic or intrinsic. Extrinsic means computing the performance of the application that uses word2vec, while intrinsic means analyzing the word relation captured by the word2vec. It is better to have extrinsic evaluation of the word2vec feature vector because looking at the intrinsic evaluation alone would not be satisfying [11]. For detailed introduction to Class-based language modeling, one can refer to [12].

We use class-based language model pp as the extrinsic evaluation metric. The classes are the clusters extracted by applying k-means clustering on the feature vectors. If a good capturing of words relationship exists in feature vectors, then similar words will belong to the same cluster.

The system architecture explained in [13] was used to get the class-based Language modeling as shown in Fig. 3.1.

3.3 Experimental Setup

We chose Iraqi dialect as our language modeling goal. This language modeling is the application on word2vec words feature vector. Thus, pp is our metric of efficient words feature vector. The test data is also an Iraqi dialect. Test data is 10% of the total corpus size, where 6% was kept blind and 4% was used for tuning.

To be able to understand how noise will affect the results, noisy data need to be provided. We made various combinations of Iraqi dialect conversations data with Modern Standard Arabic (MSA) of broadcasting formal language, which has vocabulary intersection with Iraqi and many similar contexts but nevertheless, there are a huge number of words in context that are not expected to appear in dialect phone calls. Words that are in MSA but not in Iraqi were excluded and considered as OOV.

We replicated the experimental setup of [13] shown in Fig. 2.1 for the same 5 different training sets as shown in Fig. 3.2. To understand how word feature vector size influences the results, other parameters were set fixed to window size of 5, negative sampling of 25, sub-sampling of 1e-4, and Bag of word model.

Recalling the data identification from 2.4.3:

- Set 1: Iraqi dialect [14], which is defined as low noise data since the testing set is of same style
- Set 2: Set 1 with 10% of Set 3. This Set is also considered low noise although it contains MSA context but not a big amount.
- Set 3: GALE MSA [15], which is broadcast news and reports corpus. This is a highly noisy data because it contains big amount of context that appears in broadcast news of formal language while these contexts are not expected to appear in the testing data.
- Set 4: 10 times duplicated Set 1 combined with Set 3. This is considered as noisy because it contains all of Set 3.
- Set 5: Iraqi and the whole of GALE, i.e Set 1 and Set 3.

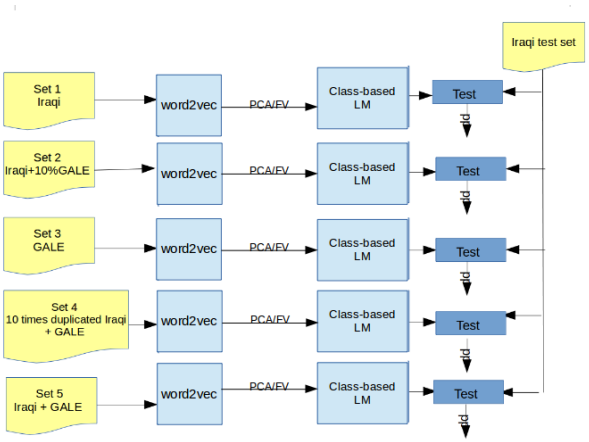


Figure 3.2.: Experimental setup flow diagram, either Features Vectors (FV) input directly to the class-based LM or PCA is applied to reduce the FV dimensionality, then input to class-based LM.

3.4 Results

Results can be analyzed in two different aspects. First, by measuring the efficiency of the words feature vector which is evaluated as to how similar words were gathered in vector space and clustered together. This is evaluated by the class-based LM pp. The second aspect is the run time delay of the application facilitating these vectors.

Table 3.1.: Perplexity reduction ratio of low and high noise data.

Training corpus	baseline vector size 350 pp	Best pp	vector size that produced best pp	Reduction ratio
set 1	137	136.4	(PCA 116 of 350)&175	0.3%
set 2	174	172.2	175	1%
set 3	500.7	466.9	(PCA 2 of 10)	6.7%
set 4	296	214.7	(PCA 2 of 10)	27.4%
set 5	372.347	291.9424	(PCA 2 of 10)	21.5%

Table 3.2.: Perplexity and clustering time using different vector size for low and high noise data.

Number	**Name**			
training corpus	number of words X vector size	k-means run time in second	vector size	pp
set 1	3233 X 2	3.6	2	160.6
set 1	3233 X 10	6.0	10	150.5
set 1	3233 X 175	21.7	175	136.2
set 1	3233 X 350	42.59	350	137
set 1	3233 X 2	3.6	(PCA 2 of 10)	149.3
set 1	3233 X 175	25.5	(PCA 175 0f 350)	136.5
set 1	3233 X 116	16.5	(PCA 116 of 350)	136.4
set 3	15651 X 2	21.4	2	473.58
set 3	15651 X 10	69.9	10	476.4
set 3	15651 X 175	212.2	175	486.45
set 3	15651 X 350	373.2	350	500.7
set 3	15651 X 2	27.7	(PCA 2 of 10)	466.94
set 3	15651 X 175	217.9	(PCA 175 0f 350)	503.38
set 3	15651 X 116	186.9	(PCA 116 of 350)	483.54

3.4.1 LM Perplexity (pp)

As explained previously, LM pp is the extrinsic metric used to measure the words feature efficiency. The base line is pp of class-based LM, where classes were produced by clustering words feature vectors of length 350. Table 3.1 shows the reduction ratio in pp. Set 3, Set 4, and Set 5 LM pp were decreased in good ratio when fetching the first two PCA components on feature vector of length 10. This means that the first two PCA components of the 10 features vector were able to represent the semantic relations between Iraqi and MSA words.

On the other hand, table 3.2 shows for low noise data as Set 1, higher pp was produced when a small feature vector is used. This means important information of the low noise data set was lost. In addition, looking back at table 3.1, we can see that no significant improvement was gained when using more than 116 PCA components of feature vector of length 350. This is shown in Fig. 3.3.

Set 4 and Set 5 are composed of GALE (MSA) but differ in that Set 5 contains one copy of Iraqi set while Set 4 has 10 duplicates of the Iraqi set. Set 4 pp was lower than Set 5 because the Iraqi contexts got highly weighted when it appeared more in the training set. Using 2 PCA components of the originally 10 length feature vector, Set 4 pp is 214.7 while Set 5 is 291.9 and thus, a reduction of 24.4% is gained.

3.4.2 Execution Time

Clustering time is one metric that we used in our implementations and it gives insights on how training and perhaps other applications could be delayed similarly. Because k-means algorithm complexity is O(number of words $*$ vector length), it is clear that reducing vector length will speed up the process as shown in Fig. 3.5.

Though, the question arises: what is a good enough vector size is needed to capture beneficial information in a clear data set? This can be answered by producing a long vector then taking the PCA that captures the most variant information.

As shown in table 3.2, for Set1, that pp was never less than 136 even when a larger vector was tested. Yet, computation time spent using vector length 350 is about three times longer

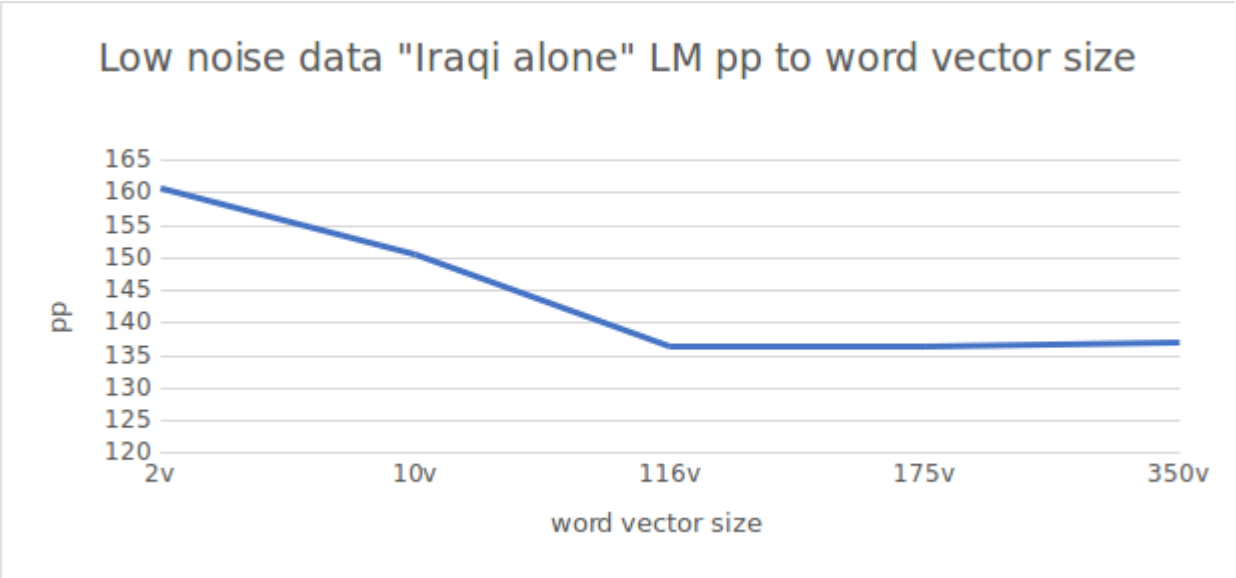


Figure 3.3.: Low noise data LM performance using different words feature vector length.

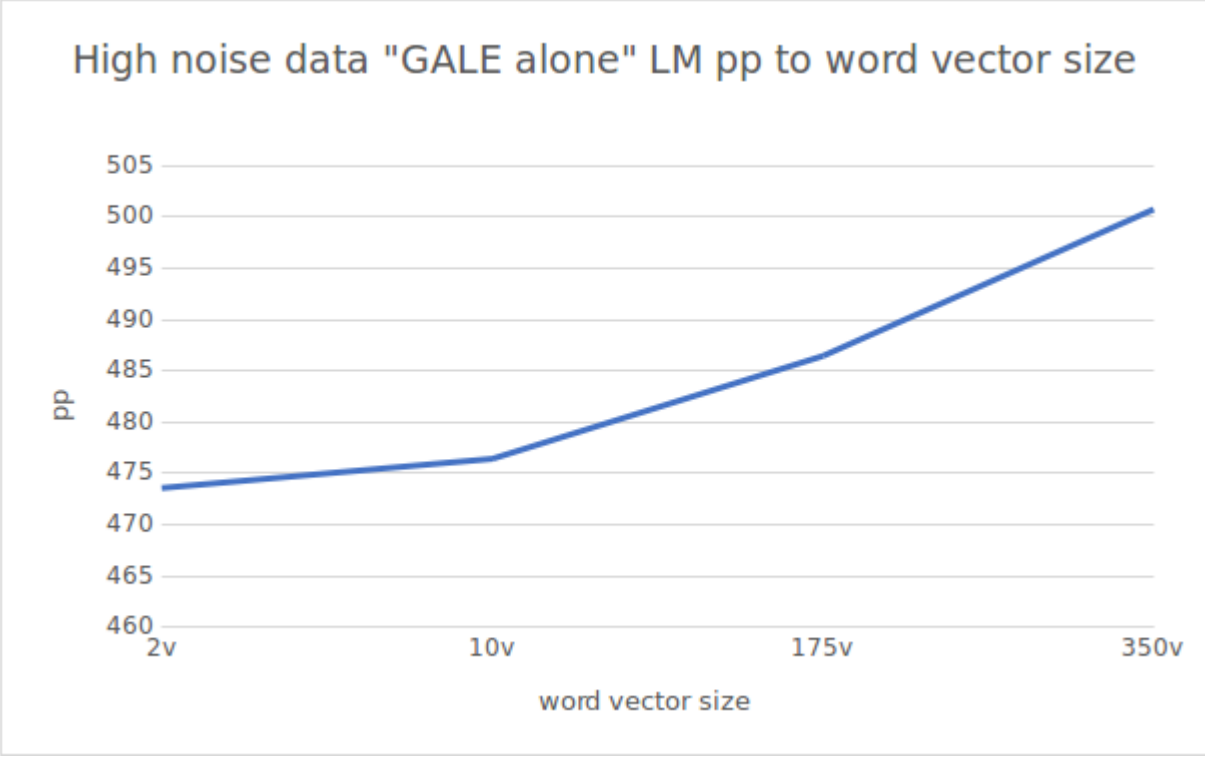


Figure 3.4.: High noise data LM performance using different words feature vector length.

than if we use 116. This is a delay with no beneficial output. The number 116 was decided after analyzing the original 350 length feature vectors and finding that about 99.7% of the

information is captured in the first third of the PCA components. See Fig. 3.6. For this reason, it is recommended to look for the minimum vector length that captures the most necessary information without wasting time and memory allocation.

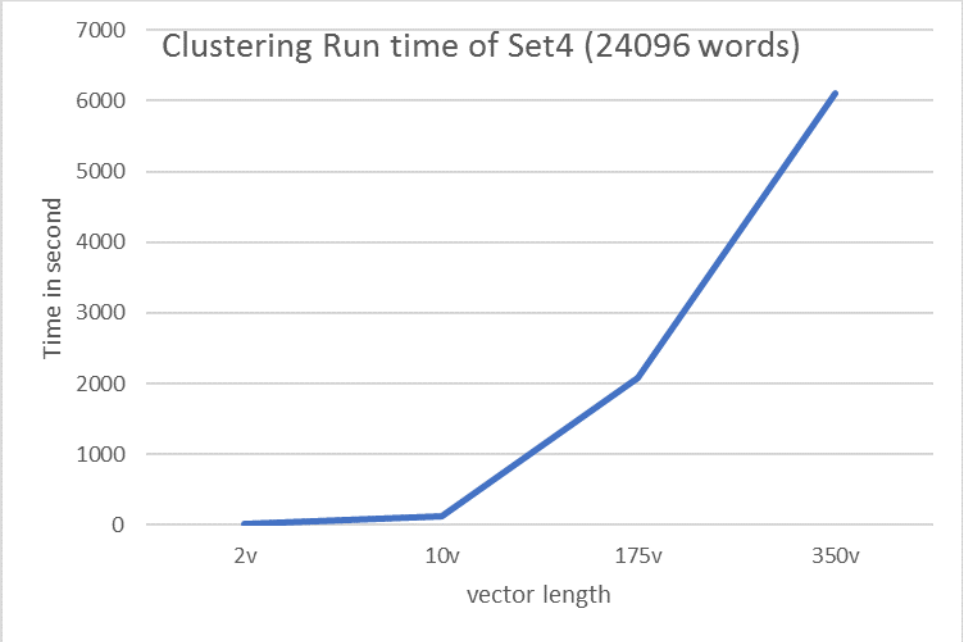


Figure 3.5.: Clustering run time for different words feature vector length.

3.5 Conclusion

By examining variations in the lengths of word feature vector of word2vec, it appears that training data size is not the only factor for deciding feature vectors lengths. Results indicate that the number of unexpected contexts (indicative of noise level) in the training set is the most important factor. The training data of low noise levels will accurately represent the information within the feature vectors with relatively long vectors. On the other hand, having noisy training data, that includes many contexts that are not expected to appear in the test data, a small feature vector length was desired. These results are also in a complete alignment of expectations with dimensionality reduction, that is noise is reduced by removing the smaller PCs, and so performance is improved.

In addition, for low noise data, there is the issue of selecting a proper vector length

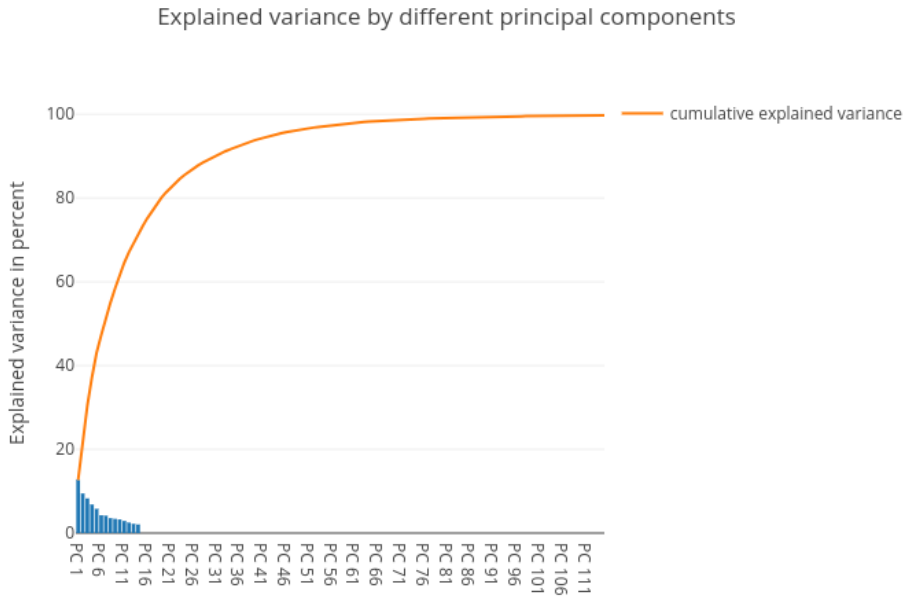


Figure 3.6.: Information in first third PCA components.

threshold. If no performance enhancement is observed, there should be no need for the vectors to be too long. A threshold value can be selected by using PCA analysis of the initial long feature vectors and retaining the components that contained within 99.8% of the covariance. For future work, conducting a comparative study using other dialects and formal languages corpora can be informative in our quest to generate an adaptive technique for word feature vector lengths.

References

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pp. 3111-3119. 2013.

[2] Seulki Bae and Youngmin Yi. Acceleration of word2vec using gpus. In Akira Hirose, Seiichi Ozawa, Kenji Doya, Kazushi Ikeda, Minho Lee, and Derong Liu, editors, Neural Information Processing, pages 269–279, Cham, 2016. Springer International Publishing.

[3] Shusen Liu, Peer-Timo Bremer, Jayaraman J Thiagarajan, Vivek Srikumar, Bei Wang,Yarden Livnat, and Valerio Pascucci. Visual exploration of semantic relationships in neural word embeddings. IEEE transactions on visualization and computer graphics 24, no. 1 (2017): 553-562.

[4] Ryan Cotterell, Adam Poliak, Benjamin Van Durme, and Jason Eisner. Explaining and generalizing skip-gram through exponential family principal component analysis. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, volume 2, pages 175–181, 2017.

[5] Chris Dyer. Notes on noise contrastive estimation and negative sampling. arXiv preprint arXiv:1410.8251 (2014)..

[6] Andrew J. Landgraf and Jeremy Bellay. word2vec skip-gram with negative sampling is a weighted logistic PCA. CoRR, abs/1705.09755, 2017.

[7] Christophe Cerisara, Pavel Krl, and Ladislav Lenc. On the effects of using word2vec representations in neural networks for dialogue act recognition. Computer Speech Language, 47:175–193, January 2018.

[8] Siwei Lai, Kang Liu, Shizhu He, and Jun Zhao. How to generate a good word embedding. IEEE Intelligent Systems, 31(6):5–14, 11 2016.

[9] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013).

[10] Sebastian Raschka. Principal component analysis in 3 simple steps. Retrieved October 20 (2015):2016 from www:http://sebastianraschka.com/Articles/2015_pca_in_3_steps.html

[11] Yulia Tsvetkov, Manaal Faruqui, and Chris Dyer. Correlation-based intrinsic evaluation of word vector representations. arXiv preprint arXiv:1606.06710 (2016)..

[12] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. Computational linguistics,18(4):467–479, 1992.

[13] Tiba Zaki Abdulhameed, Imed Zitouni, Ikhlas Abdel-Qader, and Mohamed Abusharkh. Assessing the usability of modern standard Arabic data in enhancing the language model of limited size dialect conversations. Casablanca, Moroco, December 2017. International Conference on Natural Language, Signal and Speech Processing 2017.

[14] Sydney Appen, Pty Ltd and Australia. Iraqi Arabic conversational telephone speech, transcripts LDC2006T16. 2006.

[15] Meghan Glenn and et al. GALE phase 2 Arabic broadcast conversation transcripts part 1 LDC2013T04. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

CHAPTER 4. WASF-VEC WORD EMBEDDING WASF-VEC: TOPOLOGY-BASED WORD EMBEDDING FOR MODERN STANDARD ARABIC AND IRAQI DIALECT ONTOLOGY

"Tiba Zaki Abdulhameed, Imed Zitouni, and Ikhlas Abdel-Qader. Wasf-vec: Topology-based word embedding for modern standard Arabic and Iraqi dialect ontology.ACM Trans. Asian Low-Resour. Lang. Inf. Process., 19(2), December 2019."

Word clustering is a serious challenge in low resource languages. Since words that share semantics are expected to be clustered together, it is common to use a feature vector representation generated from a distributional theory-based word embedding method. This Chapter describes the work which utilized Modern Standard Arabic (MSA) for better clustering performance of the low resource Iraqi vocabulary. We began with a new Dialect Fast Stemming Algorithm (DFSA) that utilizes the MSA data. The proposed algorithm achieved 0.85 accuracy measured by the F1 score. Then, the distributional theory-based word embedding method and a new simple, yet effective feature vector named Wasf-Vec word embedding are tested. Wasf-Vec word representation utilizes a word's topology features. The difference between Wasf-Vec and distributional theory-based word embedding is that Wasf-Vec captures relations that are not contextually based. The embedding processes is followed by an analysis of how the dialect words are clustered within other MSA words. The analysis is based on the word semantic relations that are well supported by solid linguistic theories to shed light on the strong and weak word relation representations identified by each embedding method. The analysis is handled by visualizing the feature vector in 2D space. The feature vectors of the distributional theory-based word embedding method are plotted in 2D space using the t-sne algorithm, while the Wasf-Vec feature vectors are plotted directly in 2D space. A word's nearest neighbors and the distance-histograms of the plotted words are examined. For validation purpose of the word classification used in this Chapter, the produced classes are employed in a Class-Based Language Modeling CBLM. The Wasf-Vec CBLM achieved a 7% lower perplexity (pp) than the distributional theory-based word embedding method CBLM.

This result is significant when working with low resource languages.

4.1 Introduction

The Arabic dialects, such as the Iraqi dialects, do not have enough linguistic resources for research and application development. On the other hand, Modern Standard Arabic (MSA) has been well studied, resulting in a wealth of linguistically annotated data. Since there is an intersection between the MSA and dialects, researchers should utilize the MSA to improve the dialect applications. Classifying the dialect words within the MSA words that have the same meaning, usage, or some other common features can be an important pre-processing stage for many applications. Our case study is the Iraqi dialect, but the methodology is applicable to the other dialect variants of the MSA.

Ontology is the science concerned in studying objects' properties and their relations, while semantics is the study of meaning [1]. Words are objects used to convey semantics. Thus, word ontology can be studied in a perspective of semantic theory and refers to word features and semantic relations between words. Saying two words are semantically related is not equivalent to saying that these words are synonyms, but it is true to say synonyms are semantically related because antonyms are also semantically related but in the opposite direction. Since dialects are mainly used in people's daily life conversations, this research is focusing on written words representing speech transcription. The research analyzes the semantic relations induced by word features, aiming to cluster low resource Arabic dialect words using the rich dominant MSA language.

4.1.1 Background and Problem Statement

Historically, from a linguistic view, word semantics is a deep rich philosophical topic that is concerned with word meanings. This was studied heavily by the Ancient Greeks, Indians, and Arabs. Each was influenced by their language properties [2], but nevertheless, they agreed on many common points. Unfortunately, the terminology used is quite different. Because we are studying the Arabic word ontology, we will consider some of the Arabic linguistic theories

to justify our understanding along with other Greek and Indian supporting linguistic theories.

Word Features

Natural Language Processing (NLP) treats a word as an object in the same way as other fields of research, such as signal processing, treats an image or audio signal. This object has features to be selected and extracted in order to use them in classification or for other purposes. A word's features can be extracted from the two structures. First, there is the word's topology which is the orthographic, phonological, and morphological word structure. Orthographic refers to the word's spelling, phonological refers to the word's pronunciation, and morphology is the way in which words are formed from morphemes [3]. Second, there is the word's contextual structure [4] as listed below.

1. Word topology: orthographic, phonological, and morphological word structure

   Arabic linguistics subdivide these features into:

   (a) the lexicon features which are defined as the exact meaning of a word in the dictionary. Arabic has a very large vocabulary size with a diversity of dialects across various geographical areas and often use words even from other foreign languages. In the Iraqi dialect, many Persian and Turkish words were imported during centuries of cultural interaction, such as the Persian word *Aghatee* in the Iraqi dialect used to show respect [5].

   This word's lexicon feature is also related to the acoustic features of the letters composing the word. One important characteristic of the Arabic language is having an almost direct one-to-one mapping of letter to pronunciation [6]. Theoretical studies of old Arabic linguistic scientists, such as Ibin-Ginni (941-1002) in his *Al-khasaes* book, illustrate that there is a relationship between the sound of a word and its meaning [7]. These thoughts are supported by Ancient Indian linguistic philosophers [8].

   (b) the morphological (*Sarf* is the Arabic term) features are defined by the following:

i. Rooting derivative feature that relates the word meaning to the derived root meaning, where a root has a constant lexical meaning [9];

ii. Template feature of the word that gives the template meaning of usage;

iii. Inflectional ending suffix feature of the word referring to plural-singular-dual and feminine-masculine meaning;

iv. Inflection ending suffix feature of the word referring to parsing features;

v. Inflection starting suffix feature of the word referring to feminine-masculine meaning;

vi. Inflection starting suffix feature of the word referring to the addressee - second or third person;

vii. Inflection starting suffix feature of the word referring to word negation;

viii. Inflection starting suffix feature of the word referring to future verb time.

The Arabic language follows a templatic, highly rich and complex morphological system. Its system is based on root-pattern schemes where both inflectional or/and derivation changes can be applied to the root to produce a new pattern or a new form. Different patterns relate to different syntax and semantic usage. This induces the data sparsity problem in Arabic corpora. The performances of Arabic NLP applications are highly affected by this sparsity. As an illustration, the word *give* in English may appear in five different forms. Three forms are produced by inflectional morphology that changes the word give to given, giving, and givenness, while derivation morphology produces the word gave. On the other hand, we counted the same meaning word in Arabic '> ETY'[1] أعطى in both Iraqi conversations and GALE Modern Standard Arabic (MSA) data. We were surprised by having 138 different word forms for the word *give* '> ETY' أعطى as listed in Appendix A. These were reduced to 82 after applying the morphological analyzer MADAMIRA. Arabic dialects inherit the rich morphological feature from the MSA.

Away from these Sarf template rules, people using dialects in their daily life have also

[1]Backwalter Arabic transliteration format

made some changes in some words to make them lighter in pronunciations. For example, the verb *gave* '>ETY' أعطى was replaced with '>nTY' انطى, 'ETY' عطى, and 'nTY' نطى in the Iraqi dialect. This is to reduce the difficulty in pronouncing the Hamza letter '>' أ followed by Ayn letter 'E' ع. It would be relatively heavy to pronounce the Hamza letter that produces a consonant glottal stop, followed by the voiced pharyngeal fricative Ayn letter. Both of the sounds are articulated in a very similar way, although the Ayn letter is articulated deeper down in the throat. This phenomena is named *Dissimilation* in linguistic theory [10, 11].

A word's topology features are very important factors in identifying words. Humans start to learn words initially by identifying the order of letters in a word and its shape [12]. Actually, the letters are the main identity of a word. Machine learning attempts to simulate human learning. Thus, we need to have a quantitative word feature that represents its shape. The optimal goal of machine learning in NLP research is to mimic the human cognitive system and according to [13] our brain uses functionally organized semantic memory through similarity and association between words to recognize the meaning.

2. A word's contextual structure

   This is defined by how the word is used and its collocations, or the words that accompany it. The final meaning of a sentence is understood from the composition of its words.

4.1.2 Word Relations

Now that we have illustrated a word's features, we need to identify the semantic relations that we are looking for in our analysis. Naming the semantic relations is a field of study of its own and is related to language and information retrieval. For Arabic semantic relation extraction, [14] gave a good illustration of the different approaches applied when defining the Arabic word semantic relations. In general, some studies give the taxonomy of semantic relation in a coarse-grained way. Other researchers are more domain specific and fine-grained

in their relations [15].

There are many ways to look at a word's semantic relations. Researchers define the set of relations to follow depending on the research goal and domain. In Arabic word semantic literature, Al-Gazali's (1058-1111) taxonomy is very popular. Thus, we will be looking at a combination of Al-Ghazali and Ferdinand de Saussure's (1959) classification. Al-Ghazali classified Arabic words semantic relations as the association, the part of a whole, and the inherent [16], while Ferdinand de Saussure classified word relations as paradigmatic and syntagmatic semantics relations discussed in [17]. So, relating both classifications is as follows:

1. Paradigmatic: words that are syntactically replaceable in a context [17]. This includes:

   (a) words with an association relation (Arabic term *Mutabaqa*) such as synonyms, and morphological related terms [18]. To relate all derived words to their root and other word forms, we will consider the general relation identified as the has-derived type of relation in WordNet that is used as one semantic driven feature for text classification [19]. Aristotle named this phenomenon in Greek as *paronuma* which is translated to the English language as paronyms, where a word is derived from another word and has a closely related meaning [20] such as grammar to grammarian, courageous to courage, or wisdom to wise.

   (b) words that are meronymys or related as part of a whole (Arabic term *Tadhamun*). This relation was also well approved by Al-shafie (767) who linked the relation between انسان '<nsAn' (human) to حيوان 'HywAn'(animal) to حي 'Hy' (alive) by looking at the similarities in letters order انسان '<nsAn'ending with ان 'An' is part of حيوان 'HywAn', starting with حي, 'Hy' that is part of حي 'Hy' [21]. Gilbert and the Porretan (1085) agree with Al-Shafie and illustrate the part-of-whole relation between man and human [20]. This also includes the phonological similarities [18].

2. Syntagmatic: words that are inherent (Arabic term *Mutalazima*), where words co-occur together and are often in positions near each other [17].

Nevertheless, there is a dynamic relation between the syntagmatic and paradigmatic [18].

It has been stated by [22] that distributional word-space captures both paradigmatic and syntagmatic relations and can be considered as evidences of word similarities.

In light of these theories, the main contributions in this Chapter are:

1. Development of a new feature extraction technique that addresses the high morphological properties of Arabic and dialect languages[2].

2. Achieving a higher MSA and a dialect vocabulary intersection by stemming and proposing a new dialect fast stemming algorithm (DFSA) that does not need annotated data[3].

3. Applying and analyzing distributional theory-based word embedding method on mixed MSA and dialect datasets.

4. Reducing the perplexity (pp) of the CBLM by 7%.

The rest of the Chapter is organized as follows: Section 4.2 highlights the main related work. Section 4.3 gives a description of the used data. Section 4.4 is dedicated for our approach along with the pre-processing steps. The experimental results and their analysis are listed in section 4.5. Finally, section 4.6 concludes this paper.

4.2 Related Work

To capture word features in NLP applications, the well-known word embedding techniques based on distributional theory produce vector representations of the semantic and syntactic features of the words depending on their contextual structure, i.e. word composition in sentences. On the other hand, topology features are not represented in this word embedding, and we believe the topology features have not been utilized well in extracting semantic relations, although many linguists gave evidence of the impact of these features on semantic recognition [4, 7, 12].

Work on providing a pre-trained Arabic distributional-theory based on a word embedding language model was done by [23]. This model was built by collecting Arabic text from social

[2]https://github.com/TibaZaki/Wasf-Vec
[3]https://github.com/TibaZaki/DFSA

media, Wikipedia, and other resources. A pre-process step was applied for these data sets and the LM was trained using the word2vec tool. A qualitative efficiency measurement was made by selecting a very small set of words and checking whether similar words were clustered together. In the literature of the extraction and evaluation of word analogy, i.e. relations, the original word2vec tool was tested by using a proposed set of English word analogies [24]. Similarly, [25] produced a large Arabic pre-trained feature vector from Twitter and other Arabic text resources. The efficiency was tested in the sentiment analysis process with a very limited analogy test of some words that were selected to test each word's representation. It was stated that no Arabic analogy test set was equivalent to the one produced by [24] for English or is currently available. The Google translated set from English to Arabic was not reliable and did not represent the Arabic word relations [26]. Another researcher examined the performance of the 'fasttext' tool that depends on sub-word information on an Arabic corpus, which demonstrated that it was applicable for Arabic. The evaluation was done by computing the recall and precision of test data that is publicly available and contains words that are classified as either positive or negative [27]. However, [28] recorded improvement of sentiment analysis for Arabic in a publicly available health sentiment dataset. The efficiency was evaluated by taking the five nearest neighbors of two words - good and bad. In reflecting on the usefulness of these evaluation methods, the analogy set technique was not reliable, the word classification was not dependable because words cannot always be classified as positive or negative, and the nearest neighbor technique did not give enough information.

To analyze a word's analogy, a visualization of the relations between words need to take place first. In this context, t-SNE [29, 30] is a commonly used tool in Python to reduce the feature vector to 2D and visualize the relations as distances between words. This technique was employed by [31] to compare cultural difference in word phrase usage of Korean and Japanese inter-cultural dialog. [32] focused on social network dialog to explore collective attention by choosing data from conversations during some important events and visualizing how people reacted to these events. This was done using network-based visualization of word2vec feature vectors. A similarity threshold of 0.6 for at most 20 words of a seed word, which is the most commonly used word related to the chosen event, were plotted as a network. A co-

occurrence network for all words with a frequency greater than 200 was implemented. Then, an analysis of how people were reacting to different events enabled the researchers to explore human behavior. A new method to visualize the words vectors was explored by [33]. The space dimensionality reduction was implemented in various ways using PCA, SVM+PCA, and SVM+REG to capture various words relationships. It has been stated that visualizing the word vectors in 2D space using the PCA technique will explore the relations between words with the same semantic while missing other relationships. The SVM+PCA captured parallel relations, while SVM+REG captured other relations.

Because word vector space produced by the distributional theory-based word embedding method is insensitive to the topology features, further research in the fine-tuning of the produced features vector have been taking place. The technique proposed by [34] injects morphological language-specific rules that try to either attract the related words or repel unrelated words. Another way of injecting morphological information into the word2vec feature vector was done by considering the character n-gram where the sum of these representations is considered as a word representation [35]. [36] refined the Wikipedia pre-trained word vectors through antonym repulsion and synonym attraction constraints and then complemented them with feature vectors of bag-of-word embedding trained on the ontological description text of the entities that needed to be matched. Finally, this method matched the entities of two different ontologies according to the Stable Marriage algorithm over the entities' pairwise distances. In the ontology matching application, similar words must be matched so that ontologies defined by different groups of experts can be automatically mapped. Thus, refining the learned feature vectors mainly depends on the semantic similarity relations between words. [**?**, 38] used the pre-defined synonym-antonyms relations in semantic lexicons to refine the distributional theory-based word embedding method feature vectors so that synonyms are attracted to each other. Each researcher wanted to add information to the feature vector for better word embedding.

Our approach proposes a new algorithm, the 'Wasf-Vec', that produces a new space vector representation for the word structure by employing the topology distance as will be explained later in section 4.4.2. Our approach agrees with [39], whose research was applied to English

language word pairs and showed the importance of the alphabetical word character structure to classify semantically related words and measured the words' similarities using Dice's similarity coefficient that looks for the characters' bigram similarities as the orthographic distance measure. The Wasf-Vec takes the index of the sorted vocabulary as a feature representing the character structure of words. The technique can be used in English or other languages but has higher impact on Arabic.

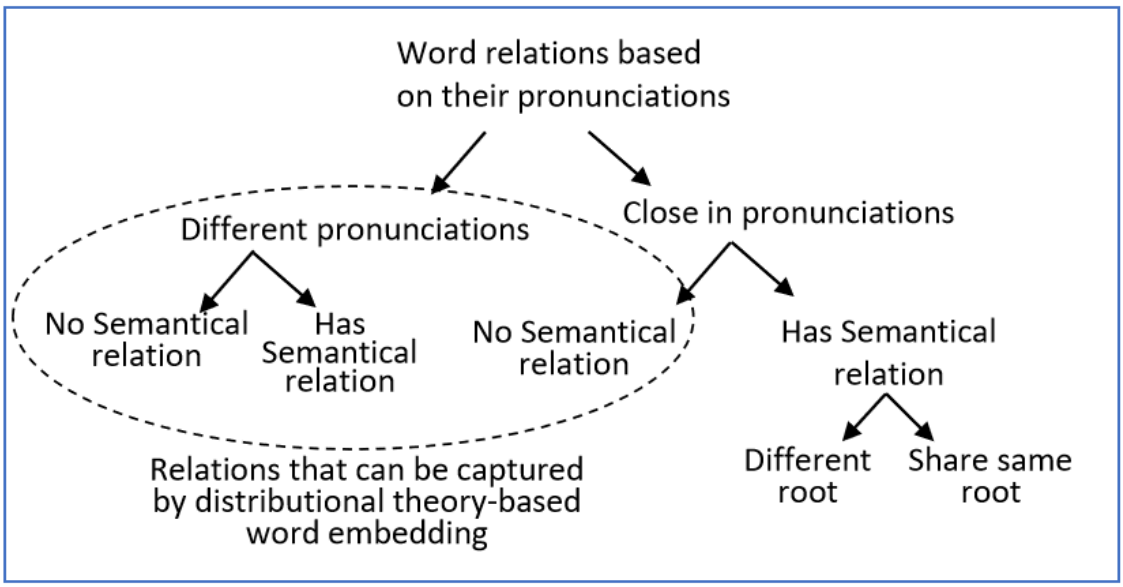


Figure 4.1.: Pronunciation similarity and semantic relations.

Fig. 4.1 shows the pronunciation relations between words. The pronunciation of an Arabic word also includes the morphological features. The distributional theory-based word embedding method can be used to capture a word's contextual features, and thus, is efficient at extracting the semantic relations between the words regardless of their pronunciation and morphological similarity degree. This word embedding covers the relations marked inside the oval area in Fig. 4.1, while Wasf-Vec is used for capturing the topology similarity degree that appears in Fig. 4.1 by following the 'close in pronunciation' branch. In addition, we propose a light dialect stemming algorithm as a pre-processing step, as explained in section 4.4.1. It is implemented on Iraqi dialect, but nevertheless, it can be implemented on other Arabic dialects with the assistance of having the MSA vocabulary list.

4.3 Datasets Description

Two corpora are used in this experimental work. The Iraqi and MSA GALE recordings are taken from the Linguistic Data Consortium (LDC).

The Iraqi Arabic Conversational Telephone Speech (LDC2006S45) [40] and their transcription (LDC2006T16) [41] are considered. This corpus contains 276 Iraqi Arabic speakers in the form of Iraqi dialect telephone conversations. The data set is subdivided as train-c1, train-c2, and devtest. Train-c1 represents one side of recorded phone conversation and train-c2 is a two-sided conversation. For our experiment, we used 90% of the corpus for training, or 199k word. The devtest is a certified standard test set according to the test process applied by the National Institute of Standards and Technology (NIST). The devtest is balanced and is 6% (102KB or about 12k words) of the total Iraqi dataset. Another 4% of the dataset was used as a tuning set.

The other data used to support the small Iraqi dialect set was the MSA GALE dataset, which contains about 1516k word of MSA broadcasts of news and reports [42, 43]. Again, the considered dataset is the transcript of the audio recording. The transcription in most cases preserves the phonology of the recorded audio. The Iraqi data size is about 10% of the MSA GALE dataset. For a balanced distributional theory-based word embedding method training, the Iraqi data set is over-sampled by duplicating the dataset 10 times, because the training is based on word and context frequencies.

4.4 Methodology

As shown in Fig. 4.2, the data is pre-processed to increase the intersections between the Iraqi dialect and the MSA. The data pre-processing techniques are illustrated in 4.4.1. Then, the processed dataset is fed to feature extraction using either a distributional theory-based word embedding method or the Wasf-Vec feature extraction technique explained in section 4.4.2. After representing the words as feature vectors, the analysis of the representation is done through visualizing the feature vectors in 2D. More details and some examples of how words were clustered with a k-means approach are found in section 4.4.4. Finally, the CBLM

is implemented to test the clustering efficiency as described in section 4.4.3.

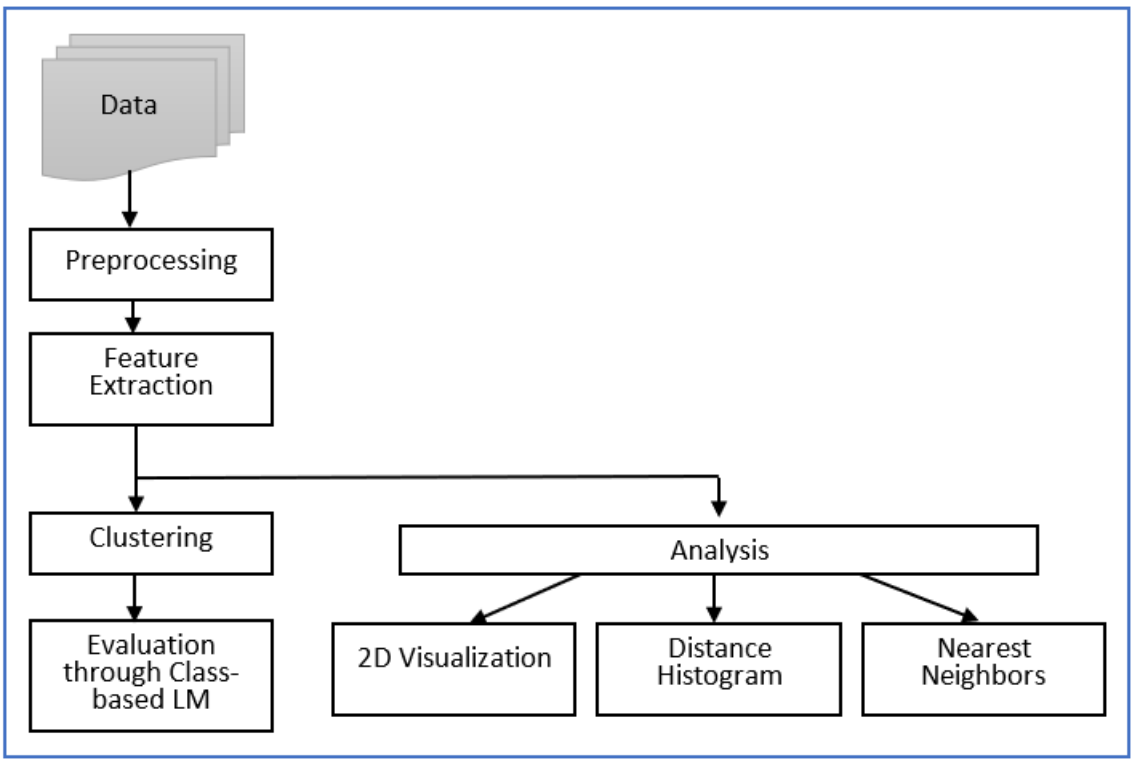


Figure 4.2.: Flow diagram of the steps for analyzing and evaluating words' features representations.

4.4.1 Data Pre-processing

Data pre-processing needs to occur to reduce the sparsity of the vocabulary set. For Arabic data, we reduce the effect of high morphology by using MADAMIRA and Iraqi stemming. Furthermore, any other foreign language written words in the transcripts are removed, and the way Hamza is written is normalized. Stemming was implemented using the MADAMIRA-release-20170403-2.1 [44]. We applied additional Iraqi stemming through our proposed Dialect Fast Stemming Algorithm (DFSA) that does not need any additional tree-bank or database. Besides, the algorithm does not need training because it depends solely on the vocabulary set and a pre-defined suffix set. The objective of the proposed algorithm is to lower the data sparsity by reducing different word forms of similar stems.

For Iraqi stemming purposes, the vocabulary set is a union of Iraqi and MSA. This is extracted from the existing corpora. The algorithm mainly reduces the Iraqi specific prefixes

Algorithm 1: Dialect Fast Stemming Algorithm

Input: $CorpusFile$,$VocabSet$,$PrefixList$, and $PostfixList$. The $PrefixList$, and $PostfixList$ should be first sorted in descending order according to the suffix length

Output: $StemedCorpusFile$.

$ListOfWords = VocabSet$;
$AscTable =$ AscendingOrderSort($ListOfWords$);
$PrefixListTemp$=$PrefixList$;
$PostfixListTemp$=$PostfixList$;
for *each word* w_i *in* $CorpusFile$, $1 \leq i \leq CorpusFile$ **do**
    $w_{new} =$ " ";
    while *length of* $w_i \geq 5$ *and* $PrefixListTemp$ *is not empty do* **do**
        get Pre_j out of $PrefixListTemp$;
        $w_{new} = w_i$-Pre_i;
        if w_{new} *in* $VocabList$ **then**
            $w_i = w_{new}$;
        end
    end
    while *length of* $w_i \geq 5$ *and* $PostfixListTemp$ *is not empty do* **do**
        get $Post_j$ out of $PostfixListTemp$;
        $w_{new} = w_i$-$Post_i$;
        if w_{new} *in* $VocabSet$ **then**
            $w_i = w_{new}$;
        end
    end
end

Figure 4.3.: Dialect fast stemming algorithm.

from a word if the remainder of the word also exists in the vocabulary set. Words that will be under-stemming consideration are of at least five letters in length, since words of less than 5 are rarely expected to be attached to prefixes. This is because most Arabic words have roots that are 3-letters long [9]. The algorithm is fast because it does not consume learning time and can be applied on MSA by defining the MSA's expected suffix set. DFSA is listed in algorithm shown in Fig. 4.3 and the same procedure is applied for post-fixes if needed. In Iraqi, we did not need to process the post-fixes further since most were captured through the MADAMIRA.

4.4.2 Features Extraction

In order to extract more information, two different methods for features extraction are to be explored:

Extracting Word Contextual Structure

The distributional theory-based word embedding methods such as the word2vec and Glove are well known techniques for feature extraction. In this Chapter, word2vec is implemented and evaluated. Word2vec captures the contextual features of words by keeping the node weights of the hidden layer of a Continuous Bag Of Words (CBOW), which is a neural network model.

Extracting Word Topology Feature Vector (Wasf-Vec Algorithm)

A new way of representing words as feature vectors is explored in this research. It is very simple, yet profound. The main idea is based on considering alphabetical order as the feature. The Wasf part of the algorithm stands for the Arabic word وصف, which means *description* in English, while, as commonly known, Vec stands for vector. So, the Wasf-Vec names the proposed algorithm shown in Fig. 4.4.

If we look at word order in a sorted vocabulary dataset, we notice clearly the importance of

Algorithm 1: Wasf-Vec: The words' topology Features Extraction

Input: $VocabSet$.

Output: $TopologicalFeatures$ matrix, a vector of 2 elements for each word as word:[feature1,feature2].

$ListOfWords = VocabSet$;
$AscTable$ = AscendingOrderSort($ListOfWords$)
for *each* w_i *in* $VocabSet$, $1<=i<=vocabSize$ **do**
  $TopologicalFeatures(w_i, 1)$ = getIndex($AscTable(w_i)$);
  $ListOfWords(i)$=ReverseLettersOrder(w_i);
end
$AscTable$ = AscendingOrderSort($ListOfWords$);
for *each* w_i *in* $VocabSet$, $1<=i<=vocabSize$ **do**
  $TopologicalFeatures(w_i, 2)$ = getIndex($AscTable(w_i)$);
end

Figure 4.4.: Wasf-Vec: The words' topology features extraction.

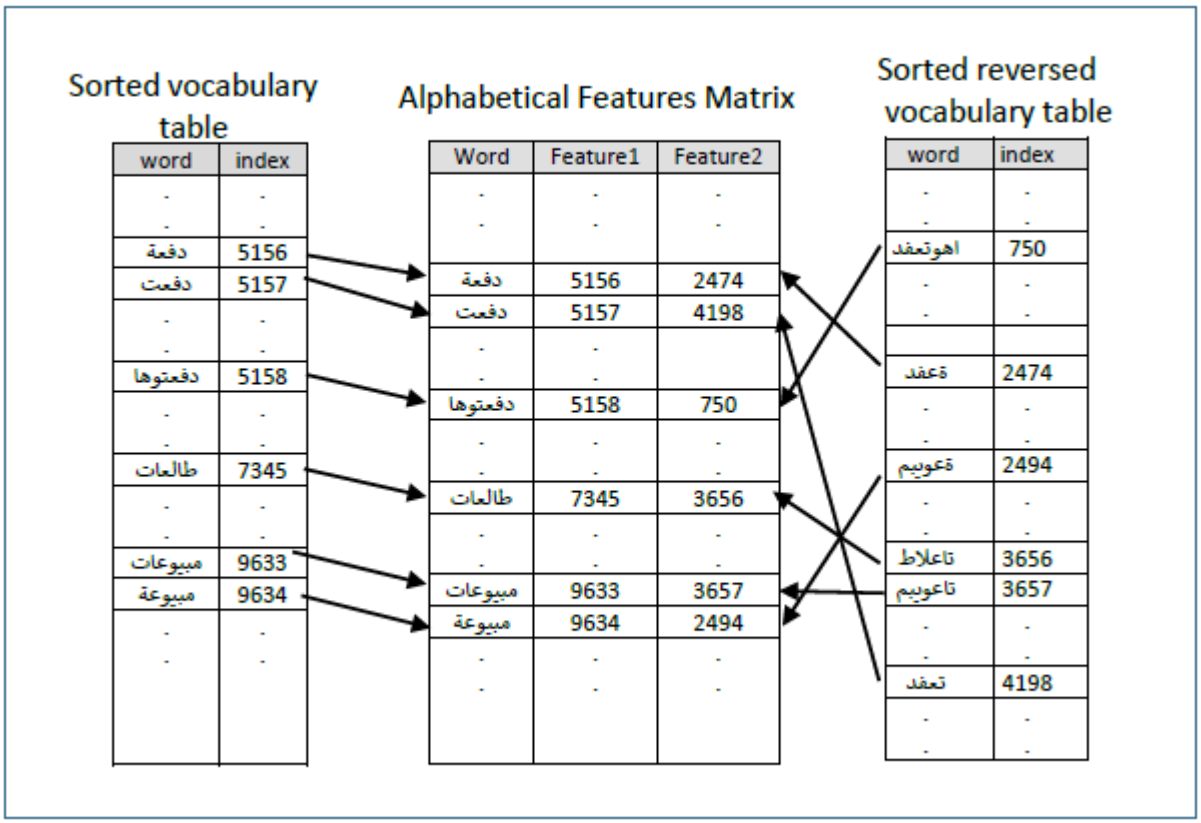


Figure 4.5.: Illustration of Wasf.

letter order. We consider the index of a word in the sorted vocabulary as its quantified lexical feature. This usage is because if we consider the one-dimensional feature to be projected at x-axis, items will be clustered in a way that related items are close together at that dimension of space. Therefore, if a word's index is the feature, we can use k-means to cluster words depending on their indexes in the sorted vocabulary list. In this way, topologically related will be gathered in the same cluster. This captures the inflectional starting, some roots, and the pattern features.

The ascending alphabetical sort of the vocabulary set will produce indexes representing the similarities at a word's beginning, but we can have another feature that represents how words are similar at the end by reversing the letter order of each word before sorting the vocabulary. This is a very important feature that captures semantic and syntactic similarities that a lexical structure enforces. For example, the feminine sound plural words will have close distances. The inflection ending suffix morphological features, as defined in section 4.1, are captured during this step. The Wasf-Vec illustration diagram is shown in Fig. 4.5.

4.4.3 Class-Based Language Modeling

This section gives a brief introduction of CBLM. To evaluate each word representation method, a CBLM is implemented and the perplexity (pp) is computed where word classes are the cluster numbers. CBLM was first introduced in Reference [45] and was implemented on word2vec classes in Reference [46]. A class-based language model would be useful when limited data resources are available. Here, the statistical information of dialect words will share the statistical information of semantically related MSA words by having the dialect words classified within the MSA words and building a CBLM. The classes are the number of the k-means clusters of the feature vectors. For comparison purposes, two different clusterings are produced, and these depend on the distributional theory-based word embedding method feature vectors and the proposed Wasf-Vec. According to the results in Reference [47], the feature vector size of the word2vec word embedding was 10 then was reduced to 2 using t-sne.

4.4.4 Analysis Method: Visualizing Feature Vectors

Because word2vec produces high dimension feature vectors, t-sne is applied to reduce the dimensionality and visuale the vectors in a 2D Cartesian plane. On the other hand, Wasf-Vec produces a 2D space vector for each word, and thus, is directly visualized.

The results of visualization and analysis of the figures are listed in the visualizing feature vector results section 4.5. The Iraqi and MSA datasets were input to build Wasf-Vec. The 10 times over-sampled Iraqi dataset and MSA GALE dataset were input to train the word2vec model. The analysis was mainly based on zooming in on arbitrary subareas of the clouds produced by plotting the feature vectors in both spaces Wasf-Vec and word2vec. Experts defined semantic relations between the words in each area. These semantic relations are represented by edges in the graphs.

The pattern fEl represents a three letter root. It carries no specific meaning except an abstract general root pattern. Words that are of this pattern retain the abstract meaning of the original lexical word [48]. On the other hand, the other Arabic patterns add semantics to the word. The root set of the Wiktionary was considered. Some other derived forms of the roots were generated and included in the analysis if they belonged to the data vocabulary set. At the same time, to have a good understanding of how words from the same root but different template relations are represented, a histogram of categorized distances is introduced. In addition to the root lexical meaning, other patterns studied were mfEwl that carries the meaning of the object where the action was done and fAEl, that carries the meaning of who did the action.

For analysis purposes, various data sets were plotted separately as listed below.

- Subset 1 contains words picked from many regions of the word2vec cloud.
- Subset 2 contains words picked from many regions of the Wasf-Vec cloud.
- Set of all words derived from *Gave* '>ETY' أعطى were plotted alone to see how these words were spread in the vector space.
- Set of some paradigmatic related words.

- Set of some syntagmatic related words.
- Set of all words that belong to the fEl and mfEwl templates.
- Set of all words that belong to the fEl and fAEl templates.

A sample of some words and their nearest neighbors that appear in words clusters are also listed to give an intuition of the clustering quality.

4.5 Results and Analysis

4.5.1 Dialect Fast Stemming Algorithm (DFSA) Accuracy

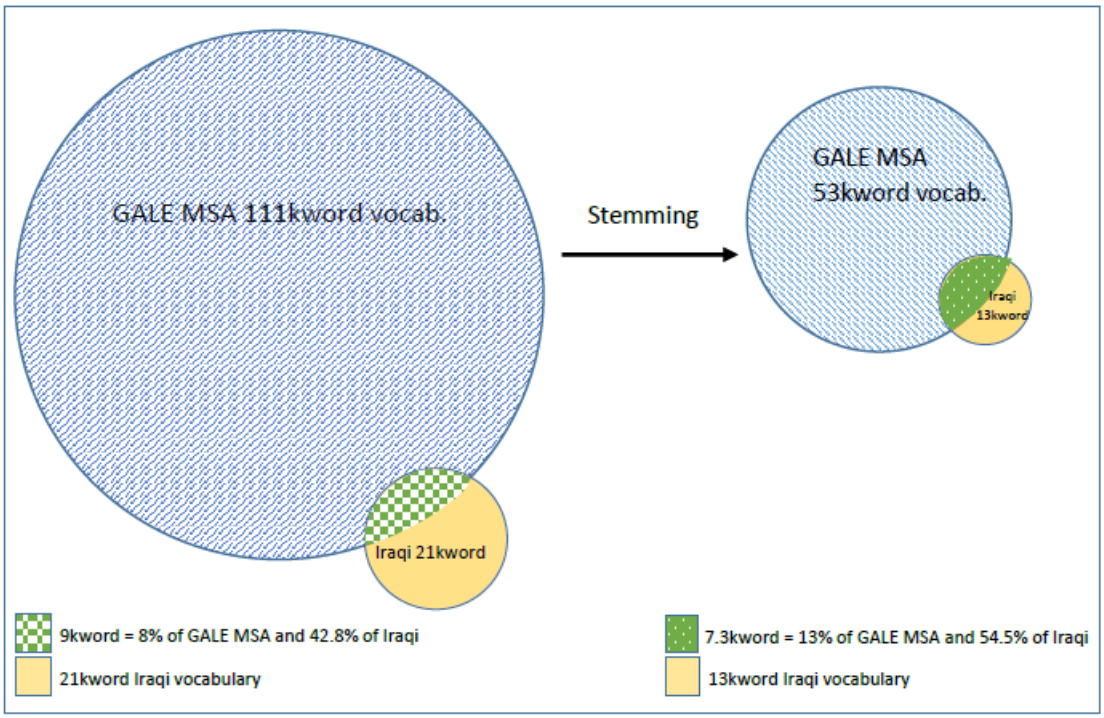


Figure 4.6.: Iraqi-MSA Vocabulary Intersections.

DFSA is an additional stemming step after applying the MSA stemming MADAMIRA. An example of stemming output of the dialect sentence:

ما ادري شاقول لك بعد شيقولون بعد ه الأوضاع الوالد مصمم إنها زينة

that was stemmed through MADAMIRA to be

ما ادري شاقول ل ّك بعد شيقولون بعد اه الّ اوضاع الّ والد مصمم ان ّها زينة

then the DFSA applied further stemming to produce

ما ادري شّ اقول ل ّك بعد شّ يقولون بعد اه الّ اوضاع الّ والد مصمم ان ّها زينة

The accuracy was computed using precision, recall and the F1 score, where:

- True positives are words that needs dialect stemming and were stemmed correctly.
- True negatives are words that do not need stemming and were left correctly.
- False negatives are words that need dialect stemming but were left out.
- False positives are words that do not need stemming but were stemmed incorrectly.

The precision is 0.94 which indicates that a very small ratio of the words was over-stemmed, while the recall is 0.78 which means there are still about 20% of the words that were under-stemmed. The F1 score can be defined as the equation 4.1 is 0.85.

$$F1 = \frac{2\,(precision * recall)}{(precision + recall)} \tag{4.1}$$

Applying both stemming techniques to the Iraqi and MSA GALE data improved the ratio of the common words to their vocabulary sets. The stemming increased the intersection ratio between Iraqi and MSA words from 42.8% to 54.5% of the Iraqi vocabulary and from 8% to 13% of the MSA GALE vocabulary, as shown in Fig. 4.6. At the same time, stemming reduced the vocabulary size of both Iraqi and MSA from 21k words to 13.4k words, and 111k words to 53k words, respectively. Although the number the intersected words of GALE MSA words and Iraqi word were reduced through stemming, the ration to the total datasets increased. Therefore, sparsity was reduced while the intersection of Iraqi words with the GALE MSA words increased which increased the relations between the two languages.

4.5.2 Visualization Result of Feature Vectors

Before starting visualization of the word relations, examples of how syntagmatic and paradigmatic relations appear in MSA and Iraqi dialect are presented in table 4.1.

Table 4.1.: Examples of how paradigmatic and syntagmatic relations between words appear in MSA and the Iraqi dialect.

relation	Paradigmatic	Syntagmatic
MSA	Used as a Rhetorical device ex1: أدلي مصدر مسؤول وقال المصدر قال ،أدلى are exchangeable synonyms*said,mentioned* ex2:من جانبه أكد الدكتور معروف البخيت وشدد البخيت في بيان شدد، أكد are exchangeable synonyms*emphasized, confirmed* ex3: ودمر الهجوم ودمر القصف الهجوم، القصف are associative word *attack, Bombing*	Grammatically correct syntax ex:- هذه حافلة تتجول على ضواحي المدن الكبرى
Iraqi	1-Dialect discourse structure contains the repetition إي إي هو أوديها إلى هذا الولد تركي إي إي هو أخذها إلى هذا الولد تركي أخذها، أوديها are associative words with similar morphology *take it, hold it.* 2- Vocabulary interchanged by other foreign words. *split* حنجيب السبلت Original MSA is المنفصلة *split unit* سنجلب وحدتالتبريد 3-Vocabulary interchanged by other variants of lighter phonemes words. ex1: عيني ها ابويا شلونك إنت *how are you* شونك ex2: ليش ثمانية ونص Original MSA word is *Half* نصف ex3: لا ألفين سعر به فرق خمسطعش ألف Original MSA word is *fifteen* خمستعشر ex4: وزعتوا شربت Original MSA word is *drinks* شروبات	1-Shorter sentences that are somehow free of constraint syntax grammar. ex: آخذها أجيبنها وسوي إل الجازجيت مالتها (no grammar used) 2- Dialect discourse use parallelism. ex1: بدبي إي أحسن والله أحسن خل ناخذ الشقة مالته أحسن نبدلها ex2: كانت شوية كانت هاي أربع ساعات شوية خطية والله ex3: هو أكو شغل حتى أكو راتب ex4: أجيب لك ألف أجيب لك ألف الفين ex5: ها ها إي هاي إذا هاي واحد واحد يشتغل عليه أه

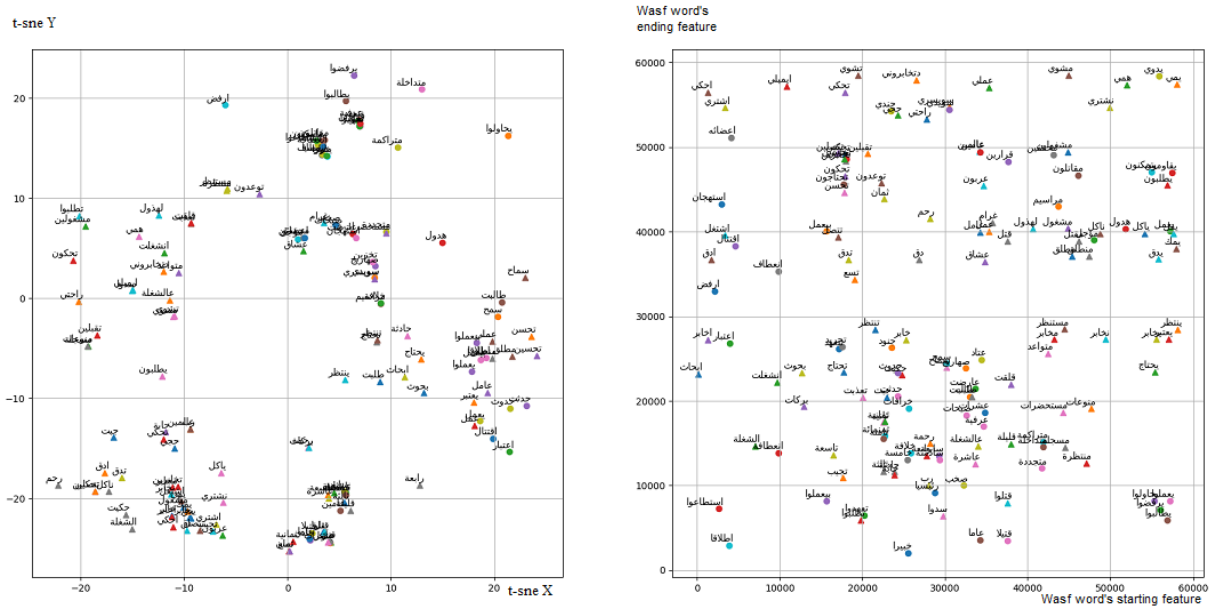


Figure 4.7.: Plots of selected set1 from different regions of word2vec cloud is shown. (a) in word2vec (b) in Wasf-Vec.

In Fig. 4.7, 4.8 words that appear in the GALE MSA vocabulary are plotted as circles, while words that belong to the Iraqi vocabulary are plotted as triangles. In the Wasf-Vec 2D

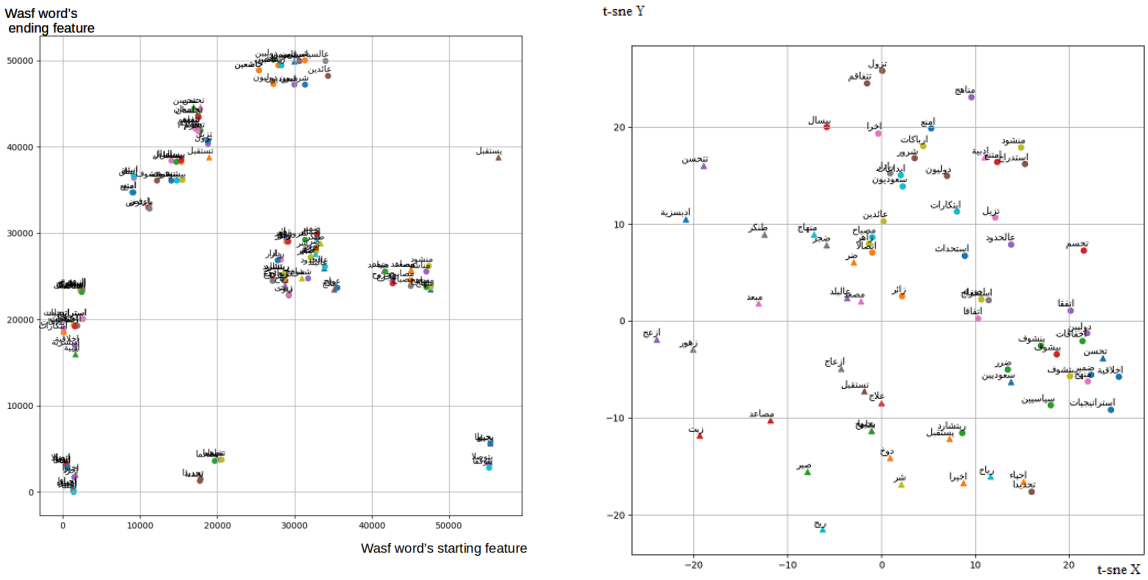


Figure 4.8.: Plots of selected set2 from different regions of the Wasf-Vec cloud. (a) in word2vec (b) in Wasf-Vec.

graph, the x-axes refers to the word topology features that are induced by the words' starting letters. On the other hand, the y-axes refers to word topology features that are induced by the words' ending letters. Since the word topology features are the indexes of sorted vocabulary, the words near to the origin of the x-axes are the words starting with first Arabic alphabetical letter 'Alef' ا, while the words starting with the last Arabic alphabetical letter 'Ya' ي appear far to the right of the x-axes. The same thing applies for the y-axes. Thus, in the top right corner are the words that start and end with 'Ya' ي.

Words plotted in the word2vec space using t-sne, On the other hand, the Wasf-vec is a 2D space, so the words are plotted by their actual feature vector values with no need to use t-sne. The scattergram Fig. 4.7, 4.8 are produced to show how same set of words is distributed/clustered in each space. Fig. 4.7 (a) is produced by zooming in on the word2vec cloud and plotting only selected words from various areas to analyze in subset 1. 4.8 (b) is produced by plotting only subset 1 selected words in the Wasf-Vec space. On the other hand, Fig. 4.8 (a) is produced by zooming in on the Wasf-Vec cloud and plotting only selected words from the various areas as subset 2. (b) is produced by plotting only subset 2 words in the word2vec space. This shows how closely related words in word2vec were spread in the

Wasf-Vec and vice versa.

To complete a better analysis, words from both subsets are plotted with predefined lines between some of the paradigmatic related words and syntagmatic related words. Thus, the length of the line represents the distance between words. Because the end goal is clustering related words, the long lines mean that the relations were not captured well and need to be attracted to each other in some manner, relation analysis should bring more insights of how to achieve this goal.

Paradigmatic Relation Analysis

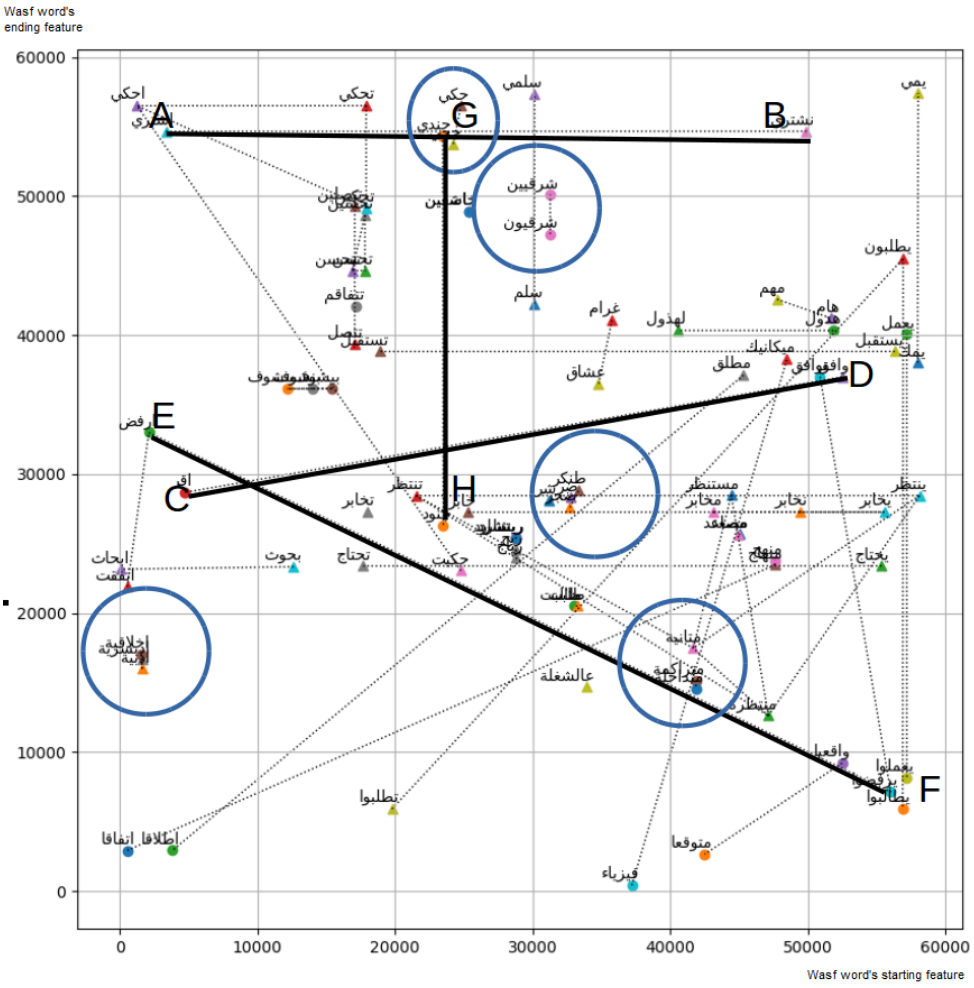


Figure 4.9.: The lines represent the distances between paradigmatic related words in Wasf-Vec.

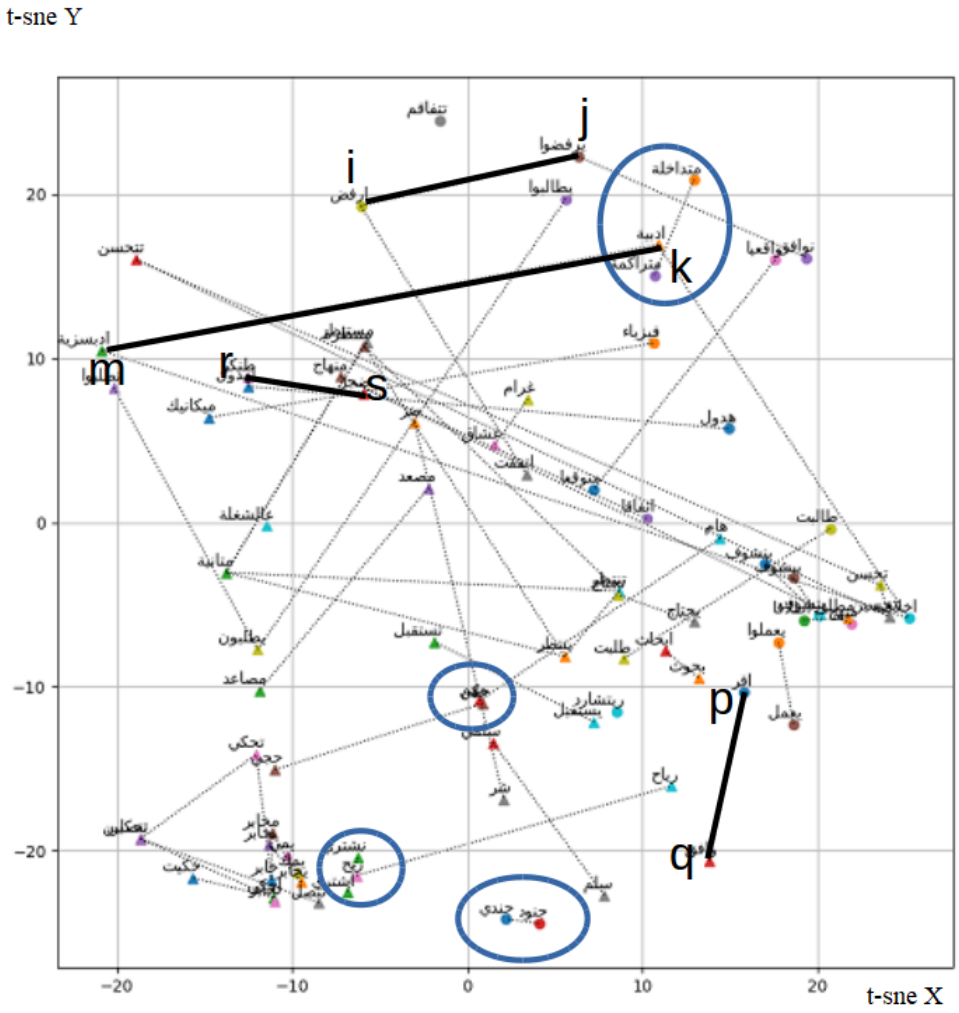


Figure 4.10.: The lines represent the distances between paradigmatic related words in word2vec.

The paradigmatic relations are very well captured by the Wasf-Vec as shown in Fig. 4.9, where words can be replaced in a sentence. Even if the edge length appears long, many straight vertical or horizontal lines can be seen. For example, the line AB connects the words *I buy* اشتري '>$tary'[4] and *we buy* نشتري 'na$tary' and line HG connects the words *officer*جندي 'jndy' and *officers* جنود 'jnuWd'. While in word2vec space, these related words appear closer to each other.

Some long-distance lines appear because related words are not from the same root, such as line CD that connects *accepted* وافق '>wAfiq' and *approved* اقر '>qir'., or they are words that share the same root but have a different prefix and post-fix added, such as line EF that

[4]Buckwalter Arabic transliteration format

connects *I refuse* ارفض '>rfuD' and *they refuse* يرفضوا 'yarfuDw'. Line ij and pq in Fig. 4.10 in word2vec space also illustrates a relatively long distances between related words.

Some Iraqi dialect words imported morphemes from the Turkish language such as the morpheme that means *without* سز 'siz' and hence, the two words *morality* and *with no morality* ادبية'>dabyap' and ادبسزية '>dabsizyap' appear close together in Fig. 4.9. While in the word2vec Fig. 4.10, the line mk shows a long distance between these words. Also, the two words for *upset*, the MSA ضجر 'Dajr' and the Iraqi dialect word طنكر 'Tankr' appear very close and have the same meaning. Although they do not belong to same root, their pronunciation is similar to each other, which agrees with Ibin-Ginni's (941-1002) theory that relates the word pronunciations to their meanings. In the word2vec space Fig. 4.10, these two words for upset appear at a greater distance than they appeared in the Waf-Vec as shown in line rs.

In Wasf-Vec, words that have the same beginning and nearly similar endings refer to plural characteristics such as *eastern people* شرقيون '$arqywn' and شرقيين '$arqyyn' are also gathered in very close positions. Moreover, words can be spelled incorrectly or differently according to the person who transcribes it, because the word is not Arabic. This applies to foreign names such as *Richard* that appears in three different forms ريتشرد 'ryt$rd' ريتشارد 'ryt$Ard ريشارد 'ry$Ard' and the word *strategy* ستراتيجي 'strAtyjy' استراتيجي '>stratyjy' that appears close to each other and in different spellings in the Wasf-Vec space. The Wasf-Vec considers all words in the vocabulary, while the word2vec considers words have some predefined frequency to enable good training. In our experiments, to be a part in the training process, a word must appear a minimum of 5 times. In the word2vec space vector, if the same word appears in different spellings, it is not recognized as the same word and is excluded from the training process. This applies to *eastern people* شرقيون '$arqywn' and شرقيين '$arqyyn'.

Both Word2vec and Wasf-Vec captured and agreed on some paradigmatic relations. For example, words that belong to the same template were gathered, such as the words *accumulated* متراكمة 'mutarAkimap' and *overlapped* متداخلة 'mutadAxilap', because they belong to the template متفاعلة 'mutafAEilap'. Given that the Arabic templates refer to how words are used and part of their meaning, the 'mutafAEilap' template means that something is interactive. This is why these words are adjacent in the word2vec space in Fig. 4.10. In the Wasf-

Vec space shown in Fig. 4.9, the words *accumulated* متراكمة 'mutarAkimap' and *overlapped* متداخلة 'mutadAxilap' also appear adjacent because they share the same template and thus have similar beginnings and endings. Dialect words that were originally MSA words but have been changed in pronunciation, such as *talk* in dialect حجي 'Hajy' and MSA حكي 'Haky', appear near each other in both Fig. 4.9, 4.10.

To compare the distances, Fig. 4.11 demonstrates (a) how word2vec captured the paradigmatic relations, while (b) how Wasf-Vec captured the same words. The Wasf-Vec histogram indicates a skewed distribution where most frequencies lie on the lower distance values. On the other hand, the word2vec histogram is spread, or has a bi-modal distribution. Actual distance values should not be used for direct comparison between Wasf-Vec and word2vec since distances belong to different vector spaces, but within each, Wasf-Vec results points to a better clustering performance. This indicates that the Wasf-Vec captured the paradigmatic relations more efficiently.

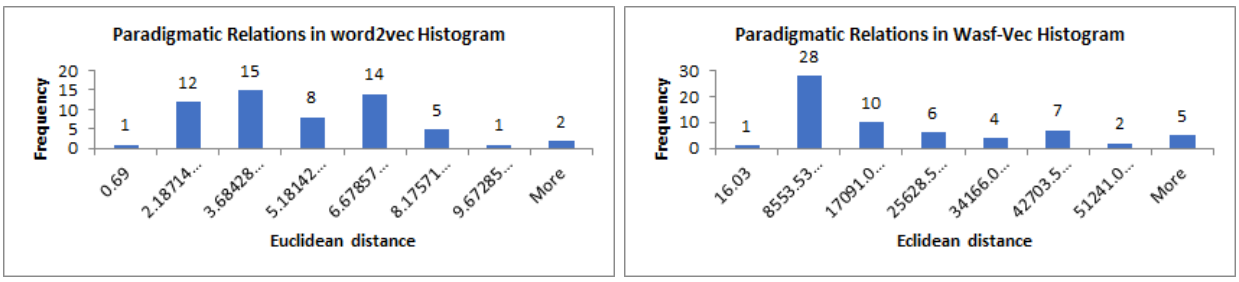


Figure 4.11.: The distances histogram of the paradigmatic related words. (a) Is word2vec space (b) Is Wasf-Vec space.

Syntagmatic Relation Analysis

Since the distributional theory-based word embedding method characterizes words by the company they keep, the syntagmatic relations are the word relations represented in the word2vec vector space. The most common case of syntagmatic relations in Arabic is defined as the *Nominal muDaf* Arabic grammar rule [9], where a word meaning is not complete until it is followed by the next word. The Wasf-Vec does not capture these relations very well as seen in Fig. 4.12 (b), while in the word2vec Fig. 4.12 (a) these words appear near each other.

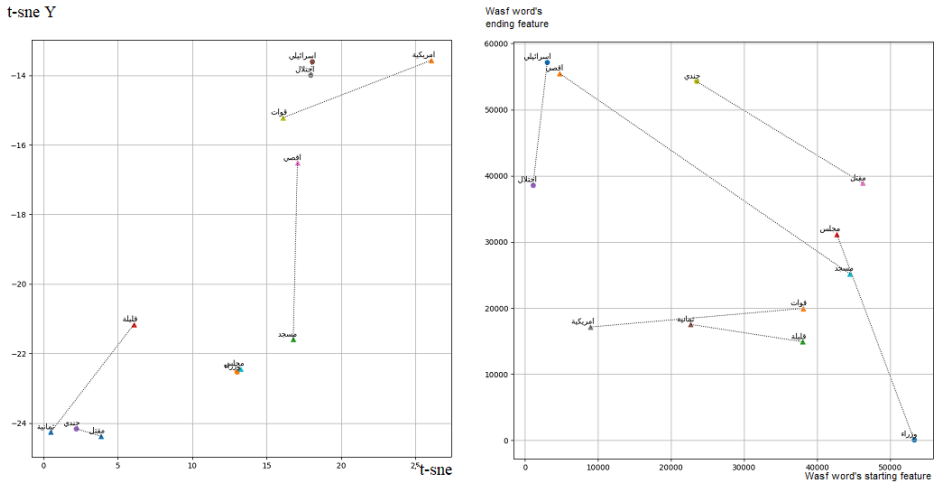


Figure 4.12.: The lines represent the distances between syntagmatic related words. (a) is word2vec space and (b) is Wasf-Vec space.

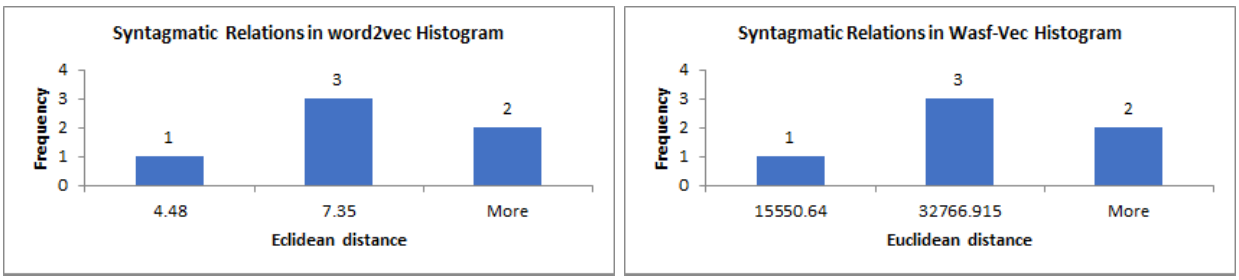


Figure 4.13.: The distances histogram of the syntagmatic relation of the words. (a) is word2vec space and (b) is Wasf-Vec space.

In terms of quantitative analysis, the distance histogram in Fig. 4.13 (a) shows how word2vec captured the syntagmatic relations, while (b) shows Wasf-Vec. Due to the small size of the syntagmatic related word sample set (12 words of 6 relations), drawing a conclusion from the histograms may not be fruitful at this point.

Same Root Word (أعطى *Gave*) Relation Analysis

Same root relation analysis can be an instance of paradigmatic relationships produced through phonology and morphology similarities. Fig. 4.14 shows the different word forms derived from '>ETY' أعطى , including the dialect specific forms. In the word2vec Fig. (a), the different word forms of the same root were spread mainly on the left side of the graph.

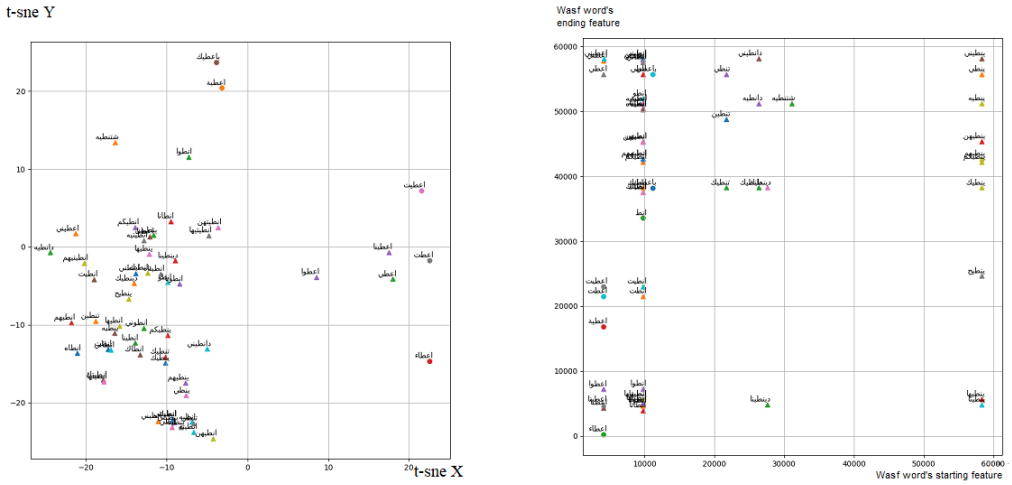


Figure 4.14.: This figure shows how different '>nTa' rooted words are spread in the vector space. (a) is word2vec space and (b) is Wasf-Vec space.

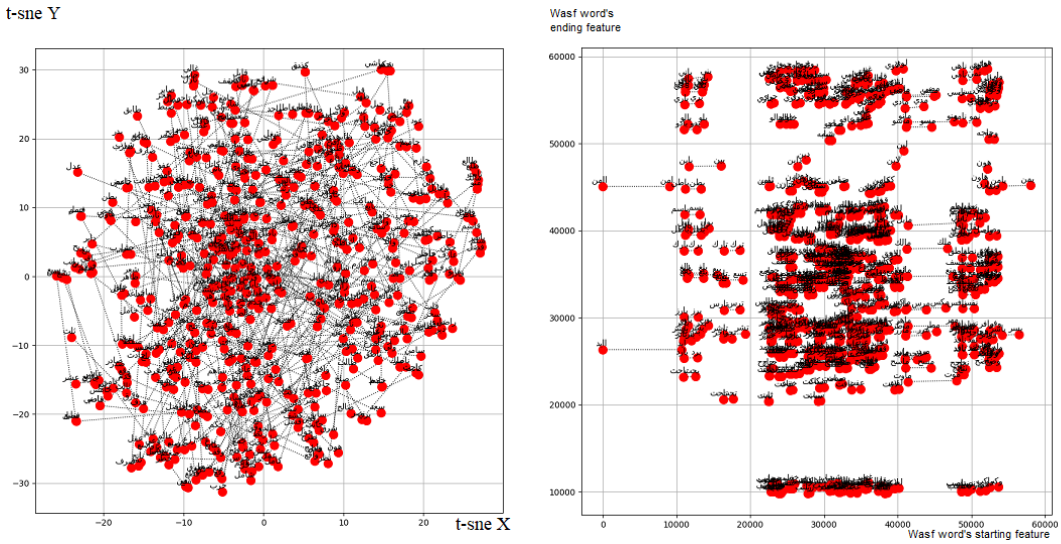


Figure 4.15.: This figure shows how words with the patterns fEl and fAEl are presented in the vector space. (a) is word2vec space, (b) is Wasf-Vec space.

While on the Wasf-Vec, Fig. (b), the present tense verbs were gathered to the right side of the Fig. and the past tense verbs appear on the left. The different pronouns attached to word endings as post-fixes cause the spreading of the words along the y-axes. The Wasf-Vec gives those different word forms a closer allocation than in the word2vec.

The fEl-fAEl, and fEl-mfEwl Patterns Relation Analysis

Root-pattern relations can be considered as syntagmatic relations because the fEL, fAEl,

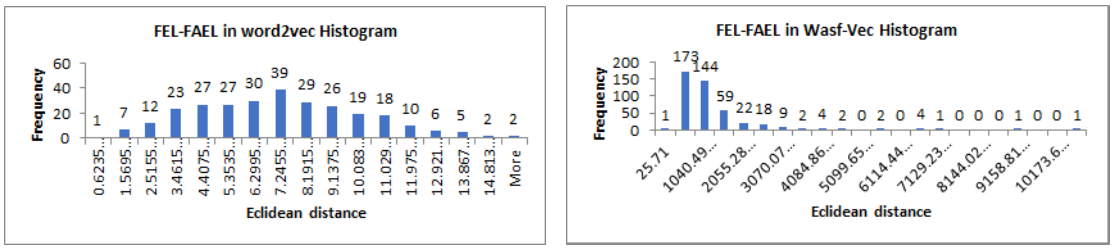


Figure 4.16.: This figure shows distance histogram of words with the patterns fEl and fAEl (a) is word2vec space, (b) is Wasf-Vec space.

and mfEwl patterns are not replaceable but have different syntax positions. To find whether there is a pattern of the word template distances, the 3-letter roots of the form fEl فعل that are defined in the Wiktionary and fAEl فاعل template that appear in our data are plotted in Fig. 4.15, where scattergram (a) is the word2vec space, and scattergram (b) is the Wasf-Vec space. Fig. 4.16 (a) and (b) are the Euclidean distance histograms. From the histograms, the Wasf-Vec has very steady distances and the distances between fEl-fAEl appear close to each other.

In Fig. 4.17, scattergrams (a) and (b) illustrate how the related words of patterns fEl-mfEwl فعل-مفعول are spread in word2vec and Wasf-Vec spaces respectively. Since all mfEwl words start with 'M', almost all appear on the right side of scattergram (b) and most share the same horizontal y-axes with the fEl form of the same word. The Wasf-Vec distances are spred in the left side, as seen in Fig. 4.18 (b). The distance histogram of the fEl-mfEwl relations in word2vec presented in Fig. 4.18 (a) can be considered as normally distributed. Thus, both Wasf-Vec and word2vec were unable to capture the fEl-mfEwl relation very well.

4.5.3 Nearest Neighbors

Now that the 2D visualization and some histograms have been illustrated, samples of the word clustering produced are shown in table 4.2 for the Wasf-Vec and 4.3 for the word2vec. Looking at the Wasf-Vec clusters sample table 4.2, the class 37 words share the same syntax rule of being masculine plural and present verb. These properties were enforced by the starting letter ي and ending letters ون. In other words, they share the paradigmatic re-

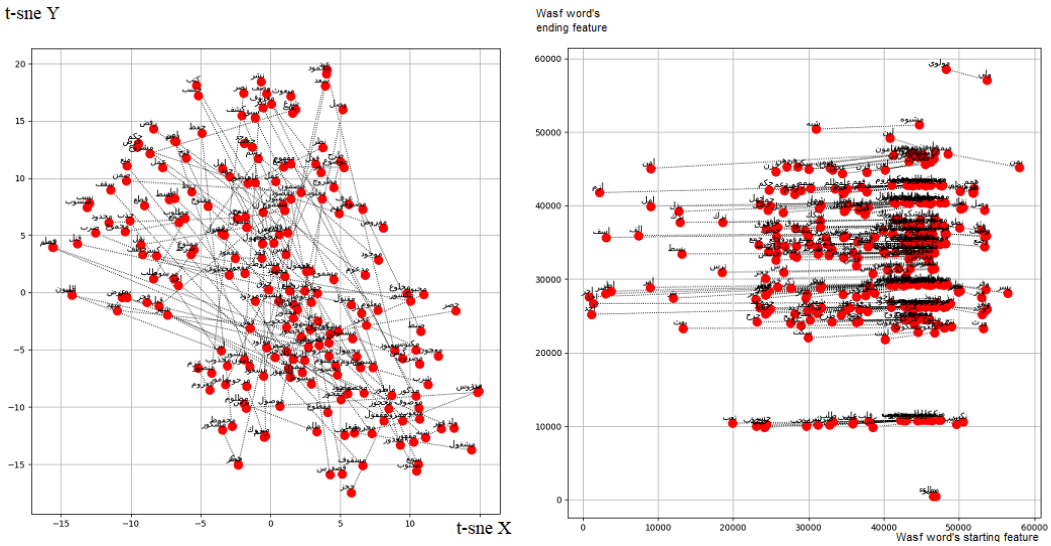


Figure 4.17.: This figure shows how words of pattern fEl and mfEwl are presented in the vector space. (a) is word2vec space, (b) is Wasf-Vec space.

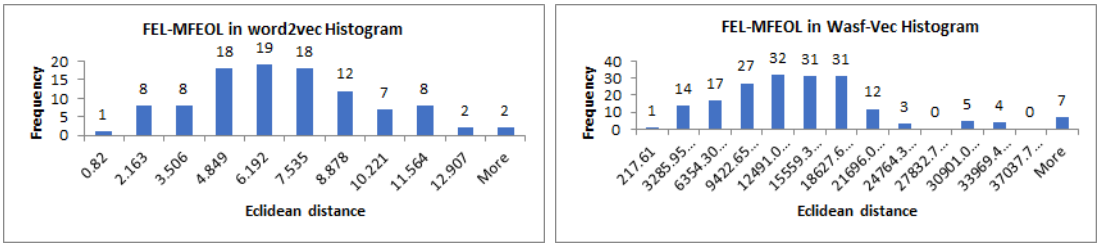


Figure 4.18.: This figure shows distance histogram of words with the patterns fEl and mfEwl (a) is word2vec space, (b) is Wasf-Vec space.

lation. Also, many words were found that have closely related meanings, such as *they fight* يقاتلون 'yuqAtilwn' *they fight one against the other* يقتتلون 'yaqtatilwn' *they resist* يقاومون 'yuqAwimwn'. Opposite words also appear, such as *they resist* يقاومون 'yuqAwimwn' *they accept* يقبلون 'yaqbalwn'. In class 22, many negative words are clustered and share the prefix ب to refer to the continuous verb and the post-fix وا to refer to the plural. Class 278 contains many names and other verbs with the prefix ف to refer to the word 'so'. Feminine plural nouns indicated by the post-fix ات are clustered in class 538. By identifying nearest neighbors, related words are clustered effectively.

The word2vec word embedding sample clusters in Table 4.3 indicate many words that can appear in same sentence. Class 37 contains many negative words that may appear in the same context. Class 23 contains *hospital* مستشفى'musta$fY' and *Upper* علوي 'Eilwy'. Upper

Table 4.2.: Sample of Wasf-Vec clustered words.

Word	Class	Translation	Word	Class	Translation	Word	Class	Translation	Word	Class	Translation
يفضلون	37	They prefer	بيشتموا	22	They are swearing	فادي	278	Fady an Arabic name		538	Comments
يفعلون	37	They do	بيشحنوا	22	They are wrangling	فارسي	278	Persian	تعليمات	538	Instructions
يفكون	37	They release	بيشككوا	22	They are doubting	فاشي	278	fascist	تفاوضات	538	Negotiations
يفهمون	37	They understand	بيشكلوا	22	They are forming	فاضي	278	empty	تعهدات	538	Promises
يقابلون	37	They meet	بيشكوا	22	They are falling under suspicion	فاقدي	278	who are losing	تعويضات	538	Compensations
يقاتلون	37	They fight	بيشنوا	22	They are triggering	فاليري	278	Falary a non Arabic name	تعيينات	538	Assignments
يقارنون	37	They compare	بيشوفوا	22	They are looking	فانتظري	278	so wait	تغيرات	538	Arbitrary Changes
يقاومون	37	They resist	بيشيلوا	22	They are raising	فبتبقي	278	so you stay	تغييرات	538	Planed Changes
يقبلون	37	They accept	بيصطفلوا	22	They taking care of their problems	فبيبتدي	278	so he starts	تفاعلات	538	inter-activities
يقتتلون	37	They fight one against the other	بيضايقوا	22	They are harassing	فبيكتفي	278	so he would be satisfied	تفاهمات	538	agreements

Table 4.3.: Sample of word2vec clustered words.

Word	Class	Translation	Word	Class	Translation	Word	Class	translation	Word	Class	translation
اتامل	37	I hope	صحيح	22	Righ	اذار	278	March	الجينز	538	Jeans
اتهامي	37	Accusing me	مالتي	22	Mine	محادثات	278	Conversations	العزايم	538	Invitations
اخلاقيا	37	Morally	نظف	22	cleanup	اتركوا	279	Leave	الغدا	538	Lunch
اذلال	37	Humiliation	يحب	22	He loves	ارشيد	279	Arabic name (Arsheed)	الكاوبوي	538	Cowboy
استبدادي	37	Tyrannical	يسوون	22	They make	استعملت	279	She used	اليم	538	Painful
استفراد	37	Standalone	يطلع	22	He is getting out	الها	279	For her	انتهينا	538	We finished
استقواء	37	Bullying	تاسعة	23	Ninth	اوميجا	279	Omega	انشق	538	Ripped
استند	37	Lean	ثمانمية	23	Eight hundred	باقول	279	I am saying	انقلب	538	Turned over
اسهام	37	Contribution	علوي	23	Upper	بتقولوا	279	You are saying	ايميلي	538	My email
اشادة	37	Praise	مستشفي	23	Hospital	بياكد	279	He is confirming	باحس	538	I feel

and hospital words are also syntagmatically related when *upper* refers to a floor. In 278, *discussions occurring in March* محادثات اذار are syntagmatically related words that are defined by the 'Mudhaf' Arabic grammar rule. But, also in 278 are *I am saying* باقول 'bAqwl' *you are saying* بتقولوا 'bitqwlw' *he is confirming* بياكد'bi>kid', which are words similar in meaning and share the prefix ب of the present tense verb to refer to continuous time. *Jeans* جينز 'jynz' *cowboy* كاوبوي 'kAwbwy' in 538 are two words of exactly the same meaning and use, so this kind of similarity would not be captured in the Wasf-Vec.

From these samples, one can conclude that the paradigmatically related words are usually clustered when the Wasf-Vec feature vector is used, while more syntagmatically related words are usually clustered together when using the word2vec feature vectors.

4.5.4 CBLM pp

The pp reduction ratio achieved by the Wasf-Vec CBLM is 7% better than the word2vec CBLM as shown in 4.4. Paradigmatic relations, which identify words that can be replaced in a sentence either for a similar meaning or for their syntax role, were represented more efficiently in the Wasf-Vec. Since CBLM basically allows same class words to share the statistical probability, then it makes sense that replaceable words share their probabilities when modeling the language. This contributes to the Wasf-Vec's pp reduction ratio over the word2vec CBLM.

Furthermore, by observing the data, the topology feature differences between the Iraqi dialect and the MSA are briefly shown in table 4.5. A thorough comparison between the MSA and Iraqi dialect can be found in Reference [9], where it has been stated that not all MSA patterns are preserved in the Iraqi dialect, and roots in Iraqi dialect may not correspond identically to the MSA roots. The table shows that the Iraqi dialect phonologically produces other variants of the MSA words that are paradigmatically related to the MSA original words. The paradigmatic relation also relates the widely used Iraqi morphological forms of words to the other MSA forms of words. Thus, the main relation between the Iraqi words and MSA words is paradigmatic, which explains why applying the Wasf-Vec and utilizing the MSA to build the Iraqi dialect LM reduces the pp by 7%.

Table 4.4.: The pp of the CBLMs.

LM	pp
Wasf-Vec class based LM	220.1413
word2vec class based LM	237.7988
Reduction Ratio	**%7.4**

4.6 Conclusions

To overcome the challenges of low resource Iraqi dialect, the first step was applying MADAMIRA stemming then using the results to perform additional dialect-specific stemming. This enlarged the vocabulary intersection and reduced the sparsity. Sparsity is a

Table 4.5.: Comparing the Iraqi to MSA in terms of topological features.

	orthographic	phonologic	morphologic
MSA	Diacritic may appear with the letters	Words pronunciations are almost one to one mapping to their letters pronunciation	Rich and variety usage of all kinds of morphological affixes.
Iraqi	Same as MSA but less need of using diacritics	Includes wider range of sounds such as /p/, /v/,/g/,/ch/. Less use of short vowel sounds on the words endings.	1-Lower use of feminine plural verb than Masculine plural verb 2-Rare use of dual format (The use of + Dual Ending ان is less than + Dual Ending ين) 3-Rare use of plural noun ending by ون

challenge because MSA is a highly morphological language, which is a property that is also passed to dialects. Indeed, both Iraqi and MSA have a large amount of words that have similar meaning but have different forms which causes the sparsity problem. Also, having many words in the dialect that are MSA words with slightly different pronunciations increased the number of word forms. The training of the distributional theory-based word embedding method is significantly impacted by this sparsity. To overcome this, the Wasf-Vec was developed to reduce the sparsity issue. The Wasf-Vec is a word embedding system that takes advantage of words topological features. Allowing words to be close together based on the topology distance space resulted in assigning these words to the same statistical class in the CBLM. This significantly reduced the sparsity problem.

Two types of word representations for the Iraqi dialect and MSA were used in order to develop word clusters. By applying the distributional theory-based word embedding method and the Wasf-Vec, two types of clustering were produced. The clustering results show that Wasf-Vec represented the paradigmatic relationships efficiently and proved to be more reasonable for sharing probabilities through CBLM.

In future investigations, it would be worthy to investigate a system in which both feature vectors are used for classifying MSA and dialect words. Verifying Wasf-Vec's abilities on other data sets to allow performance quantification via word pp to develop a better understanding of

the new technique and its advantages. Also, it would be good to investigate deriving the Arabic analogy set from various patterns that share the same root and/or nominal, *muDaf*, phrased words to test the efficiency of the distributional theory-based word embedding method.

References

[1] Don Lee Fred Nilsen. Semantic theory: a linguistic perspective / Don L. F. Nilsen, Alleen Pace Nilsen.1975.

[2] Wout Van Bekkum, Jan Houben, Ineke Sluiter, and Kees Versteegh. The emergence of semantics in four linguistic traditions: Hebrew, Sanskrit, Greek, Arabic. 1997.

[3] Zakaria Mohamed Kurdi. Natural language processing and computational linguistics: speech, morphology and syntax. Vol. 1. John Wiley & Sons, 2016.

[4] Ahmed Mukhtar Umer. علم الدلالة [Ilm al-dilalah]. Alam Alkotob, Cairo, 5th. edition,1998.

[5] Zaki Abdulhameed Alhabba. مُفرَدات فَارسيَة في بَغداديات عزيز حجية[Persian vocabulary inBaghdadiat Aziz Hejiyah]. Arab Encyclopedia House, 2002.

[6] Fadi Biadsy, Julia Hirschberg, and Nizar Habash. Spoken Arabic dialect identification using phonotactic modeling. In Proceedings of the eacl 2009 workshop on computational approaches to semitic languages, pages 53–61. Association for Computational Linguistics,2009.

[7] Malika Sadi. الدلالة و جدل اللفظ و المعنى [Significance and dialectic of the word relation tomeaning].Oudnad, 61, 2011.

[8] Aryeh Levin. Arabic linguistic thought and dialectology. JSAI, 1, 1998.

[9] Salih J Altoma. The problem of diglossia in Arabic: a comparative study of classical and Iraqi Arabic. Harvard Middle Eastern monographs; 21. Distributed for the Center for Middle Eastern Studies of Harvard University by Harvard University Press, 1969.

[10] Ahmed Mukhtar Umer. اسس علم اللغة [A Basic Introduction to the Science of Language]. Alam Alkotob, 8th edition, 1998.

[11] Mario Pei. Invitation to linguistics: a basic introduction to the science of language. Doubleday, 1st edition, 1965.

[12] Ram Frost and Leonard Katz. Orthography, phonology, morphology, and meaning: An overview. Advances in Psychology, 94(C):1–8, 1992.

[13] Kutas and Federmeier. Electrophysiology reveals semantic memory use in language comprehension.Trends in cognitive sciences, 4(12), 2000.

[14] Wiem Lahbib, Ibrahim Bounhas, Bilel Elayeb, Fabrice Evrard, and Yahya Slimani. Ahybrid approach for Arabic semantic relation extraction. In FLAIRS Conference, 2013.

[15] Siegfried Handschuh, Vivian S Silva, Manuela Hürliman, André Freitas, and Brian Davis. Semantic relation classification: task formalisation and refinement. CogAlex-V@ COLING 2016, 2016.

[16] Sarah Haydosi. الدرس الدلالي عند الاصوليين، أنموذجا الشافعي [The semantic lesson of fundamentalists, Case study-Alshafie]. Master's thesis, Algeria, 2016. Larbi ben Mahidi OumEl Bouaqhi.

[17] Christopher SG Khoo and Jin-Cheon Na. Semantic relations in information science. Annual review of information science and technology, 40(1):157–228, 2006.

[18] Barbara Johnstone.Repetition in Arabic discourse paradigms, syntagms, and the ecology of language / Barbara Johnstone. Pragmatics beyond ; new ser., 18. John Benjamins Pub., Amsterdam ; Philadelphia, 1991.

[19] Suhad A Yousif, Venus W Samawi, Islam Elkabani, and Rached Zantout. Enhancement of Arabic Text Classification Using Semantic Relations of Arabic WordNet Text Classifi-cation based on Semantic Relations View project Failure Recovery in Distributed Storage Systems View project.Journal of Computer Science, 11(3):498–509, 2015.

[20] Barbara Cassin, editor.Paronym, Derivatively Named, Cognate Word. Princeton University Press, 2013.

[21] Muhamed Abu Zahrah. الشافعي حياته وعصره واراؤه الفقهية [al-Shafie Hayatuh wa asruh waarauh al-faqhieah]. Dar al-fiker al-araby, cairo, 2nd edition, 1978.

[22] Magnus Sahlgren.The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. PhD thesis, Stockholm University, US-AB, Sweden, 2006.

[23] Abu Bakr Soliman, Kareem Eissa, and Samhaa R. El-Beltagy. Aravec: A set of Arabic word embedding models for use in arabic nlp. Procedia Computer Science, 117:256–265,2017.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. 2013.

[25] A. Aziz Altowayan and Lixin Tao. Word embeddings for Arabic sentiment analysis. In 2016 IEEE International Conference on Big Data (Big Data), pages 3820–3825. IEEE,2016.

[26] Mohamed A Zahran, Ahmed Magooda, Ashraf Y Mahgoub, Hazem Raafat, Mohsen Rashwan, and Amir Atyia. Word representations in vector space and their applications for Arabic. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 430–443. Springer, Cham, 2015.

[27] A. Aziz Altowayan and Ashraf Elnagar. Improving arabic sentiment analysis with sentiment-specific embeddings, pages 4314–4320. IEEE, December 2017.

[28] Abdulaziz M Alayba, Vasile Palade, Matthew England, and Rahat Iqbal. Improving sentiment analysis in Arabic using word representation. In 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), pages 13–18.IEEE, 2018.

[29] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms.The Journal of Machine Learning Research, 15(1):3221–3245, 2014.

[30] Laurens Van Der Maaten and G Hinton. Visualizing data using t-sne. Journal Of MachineLearning Research, 9:2579–2605, 2008.

[31] Heeryon Cho and Sang Min Yoon. Issues in visualizing intercultural dialogue using word2vec and t-sne, pages 149–150. IEEE, September 2017.

[32] Kazutoshi Sasahara. Visualizing collective attention using association networks. New Generation Computing, 34(4):323–340, October 2016.

[33] Shusen Liu, Peer-Timo Bremer, Jayaraman J. Thiagarajan, Vivek Srikumar, Bei Wang,Yarden Livnat, and Valerio Pascucci. Visual exploration of semantic relationships in neural word embeddings. IEEE Transactions on Visualization and Computer Graphics,24(1):553–562, 1 (2017).

[34] Ivan Vulić, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen. Morph-fitting: Fine-tuning word vector spaces with simple language-specificr ules. In ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), volume 1, pages 56–68. Association for Computational Linguistics (ACL), 2017.

[35] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information.Transactions of the Association for Computational Linguistics, 5, 2017.

[36] Prodromos Kolyvakis, Alexandros Kalousis, and Dimitris Kiritsis. Deep alignment: Un-supervised ontology matching with refined word vectors. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies, Volume 1 (Long Papers), volume 1, pages 787–798, 2018.

[37] Nikola Mrkšić, Diarmuid O Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting word vectors to linguistic constraints. In 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference. Association for Computational Linguistics (ACL), 2016.

[38] M. Faruqui, J. Dodge, S.K. Jauhar, C. Dyer, E. Hovy, and N.A. Smith. Retrofitting word vectors to semantic lexicons. In NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pages 1606–1615. Association for Computational Linguistics (ACL), 2015.

[39] George W. Adamson and Jillian Boreham. The use of an association measure based on character structure to identify semantically related pairs of words and document titles. Information Storage and Retrieval, 1974.

[40] Sydney Appen, Pty Ltd and Australia. Iraqi Arabic conversational telephone speech LDC2006S45. Web Download. Philadelphia: Linguistic Data Consortium, 2006.

[41] Sydney Appen, Pty Ltd and Australia. Iraqi Arabic conversational telephone speech, transcripts LDC2006T16. 2006.

[42] Meghan Glenn and et al. GALE phase 2 Arabic broadcast conversation transcripts part 2 LDC2013T17. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

[43] Meghan Glenn and et al. GALE phase 2 Arabic broadcast conversation transcripts part 1 LDC2013T04. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

[44] Arfath Pasha, Mohamed Al-Badrashiny, Mona T. Diab, Ahmed El Kholy, Ramy Eskan-der, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In LREC, 2014.

[45] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. Computational linguistics,18(4):467–479, 1992.

[46] Tiba Zaki Abdulhameed, Imed Zitouni, Ikhlas Abdel-Qader, and Mohamed Abusharkh.Assessing the usability of modern standard Arabic data in enhancing the language model of limited size dialect conversations. Casablanca, Moroco, December 2017. InternationalConference on Natural Language, Signal and Speech Processing 2017.

[47] Tiba Zaki Abdulhameed, Imed Zitouni, and Ikhlas Abdel-Qader. Enhancement of the word2vec class-based language modeling by optimizing the features vector using pca. In 2018 IEEE International Conference on Electro/Information Technology (EIT), volume 2018-, pages 0866–0870. IEEE, 2018.

[48] Mahmud Fahmi Hijazi.اسس علم اللغة العربية ['Usus 'ilm al-lugat al-'arabiyya]. دار الثقافة [Dar at-Taqafa], Cairo, 1978.

Because the daily spoken dialects are still considered to be low-resource languages, Arabic NLP uses MSA. Since these dialects, such as Iraqi, are mainly derived from MSA, one should aim to benefit from their properties' intersections. Language properties are vocabulary and linguistic features such as morphology and phoneme sets. The goal of this thesis is to build an LM for dialect conversations to be used in a speech recognition system. Not only are the challenges of MSA processing passed to the derived dialects, but also, conversational type of communication increases the challenge level due to the differences in word usage and utterance length.

Developing the LM is a fundamental process to build many NLP applications. In this work, we opted to model the problem as a domain adaptation. That is MSA and dialect are being viewed as the same language but different domain. The differences come in two major axes: difference in speech style (such as in news vs. conversation), and difference in vocabulary set. Interpolated LM and class-based LM are selected to address the domain adaptation problem.

It has been proven that word embedding produces good unsupervised word feature vector representation. The need for unsupervised classification stems from the fact that dialects are low-resource languages, and researchers do not have a pre-defined annotated vocabulary set. The clusters are used for building the class-based LM, and those clusters are the media that carry the transferred information between the two languages. Thus, efficient clustering sustains better cross-language information transformation.

The distributional-based word embedding applied in this thesis is NN-based using word2vec, where for each word a semantic and syntactic feature vector were extracted. Investigations also included a selection criterion for best feature vector length that represents a word. Principal component analysis (PCA) was used at the feature vector level and found that small size feature vectors give better representation of words if the NN word embedding is trained on different domain corpora. This eliminated some of the noise effect related to having words with context that were not expected to occur in test data. Syntax and semantic features based

on word context usage are not always applicable or even consistent. For instance, broadcast news and report contexts are very different from phone conversation contexts.

Moreover, because Wasf-vec is a Topology-based word embedding, it overcomes the limitations of having relatively small data for training, and essentially solves the sparsity problem caused by the high morphology property of the Arabic language.

An analysis of the weaknesses and strengths of both word embeddings in terms of syntagmatic and paradigmatic relations is presented with a solid linguistic theoretical base. Paradigmatic relates replaceable words, while syntagmatic relates words in some context. Our analysis shows that the paradigmatic relation is the most important relation when considering cross-language word relation of dialect and MSA. That is why the LM that was based on classes produced from clustering words in Wasf-Vec outperformed the LM that was based on classes produced from clustering words in distributional-based word embedding.

Deploying the additional new stemming algorithm, DFSA, increased the vocabulary intersection between the two domains. In addition, the assumption that dialect and MSA are two different but intersected domains answers the question regarding why the interpolated LMs of both domains resulted improvements.

The LM efficiency was calculated by its perplexity metric, but other methods to evaluate the LM can be used, such as implementing it in an NLP application and calculating the enhancement achieved. However, the lack of access to the ASR system to test the proposed model limited this work to perplexity metric, and therefore, a reduction in WER was not presented.

The importance of using Wasf-vec, a topology-based word embedding method, is to reduce sparsity that could not be achieved using other word embedding methods. It is important to locate different forms of words in order to share statistical history that a class-based language model will use. Thus, taking advantage of word topological features allows the recognition of different word forms in one cluster and assists in identifying the most common language utterance spoken. A speech recognition system can be built using the new language model to develop a lower WER in the future.

5.1 Contribution

In this chapter, we give an overview of this dissertation contribution and few suggestions to future work.

- **Adapting MSA LM to a dialect LM**. We defined the problem of utilizing MSA data to develop a dialect-LM as a domain adaptation. This definition leads to interpolating both of the languages' LMs. It is illustrated in this work how both domain adaptation is successful and is one way of utilizing the MSA data to support developing the dialect LM.

- **Class-based language modeling**. To address the difficulties associated with limited dialect data, we used classes that were obtained via clustering the word embeddings of MSA-dialect words. Clustering the dialect words within the MSA words allowed the statistical information transformation of the rich MSA data to the dialect.

- **Specifying noise in the feature vector produced by NN word embedding**. Distributional-based word embedding of multiple domain-training corpora is defined as context that appears in one domain but not in the other. This noise is reduced using PCA.

- **Words ontology analysis**. Addressing the definition of word ontology using word contextual and topological features and its paradigmatic and syntagmatic related words. This definition allowed us to identify the information that were underrepresented in the word embedding.

- **Intrinsic evaluation of word embedding**. A deep look at the word representation through visualization and distance histogram of selected paradigmatic and syntagmatic related words shows the intrinsic evaluation of the representation in addition to the extrinsic evaluation that was computed through LM pp.

- **Iraqi- MSA analogy analysis from information technology insight**. This is new development that allowed intrinsic information of the Iraqi dialect word embeddings to

be investigated and tied to its counterpart in MSA. Designed and Developed Dialect Fast Stemming Algorithm (DFSA). This framework has several elements of novelty such as the stemming algorithm which can be adopted for any dialect that has a corresponding corpus, such MSA.

- **Wasf-vec word embedding framework**. This algorithm is based on topological word features that solved the Arabic high morphology issue and can be applied to other languages as well. The algorithm is simple yet very effective and is efficient in cross language information transformation by capturing relations that were missing in distributional word embedding. The data sparsity caused the high morphological property of Arabic is highly reduced by clustering various forms of same rooted words in same cluster and consider the class-based LM.

5.2 Future Work

The findings of this research will allow for building a new structure for a probabilistic NN LM to adjust the weights of the NN of distributional-based word embedding, which are the semantic and syntactic features, for more efficient words representation. The adjustment aims to integrate the proposed topological features with the semantic and syntactic features in one probabilistic NN LM.

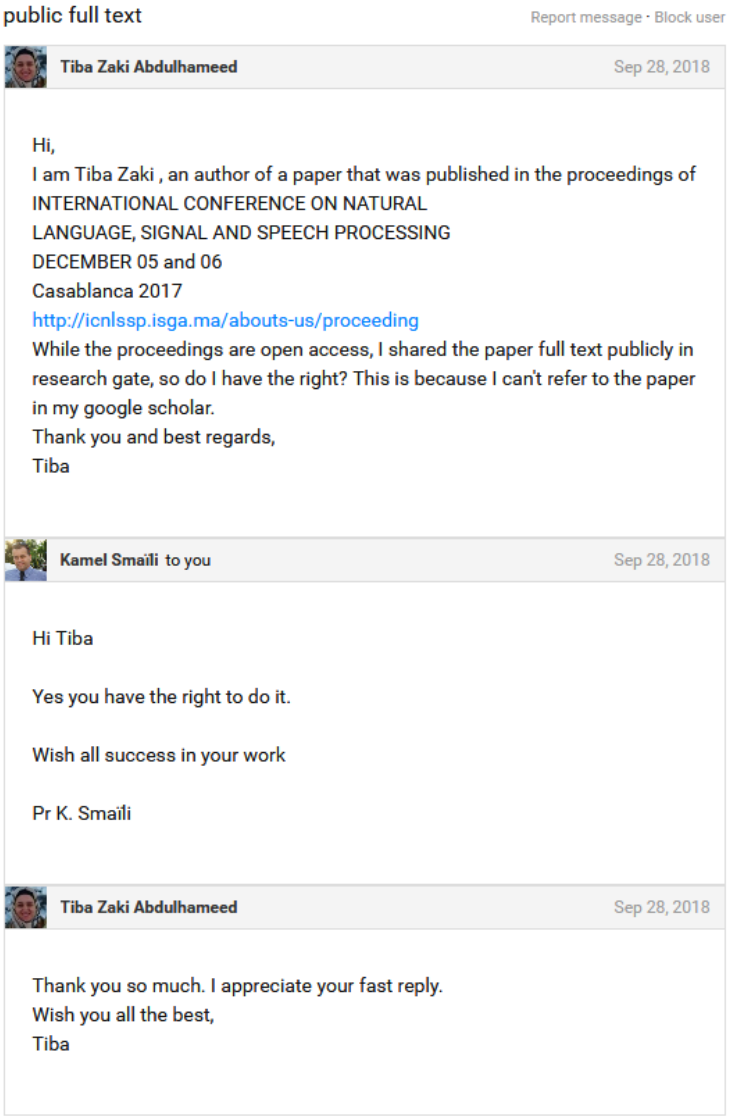
Further, the work allows for applying the resultant LM to a Kaldi ASR recipe for Iraqi phone conversation data to evaluate the proposed LM with WER metric. This will give a more solid validation of the LM to be used in actual applications.

In addition, the Wasf-Vec can be tested in other NLP applications such as dialect sentiment analysis and translation. We expect that capturing the paradigmatic relations using Wasf-Vec will cause improving the performance of these applications. This can be done by concatenating topological features with the semantic and syntactic features and feeding them to a DNN classification system.

A Different Forms of the Word Give

انطاه حينطيك ينطيني أنطاه ينطيك دينطينا انطيك وأنطيك دأنطيه وأنطاه أعطاها
وينطيني وانطيك تنطين تنطيه تنطينا وانطيت انطيت حينطيه إنطيه إنطونا
حانطيه إنطي دانطيني أنطيك انطيتيه إنطيها وتنطي انطيتيهم تنطيهم إنطيتني
إنطوه وأعطى أنطيه إنطاني إنطاك وانطوه أعطيه انطانا وانطاني وانطيتها
ينطيها أعطيتك ينطيه وتنطيني تنطي انطاك دينطيك انطيهم انطيتيها أعطينا
انطته تنطيها وإعطاء انطتها أنطتني انطينا ينطينا وأعطاه انطيته انطيناه
وأعطينا أعطونا وأنطي انطتك انطوك وأعطت انطوني اعطنا أعطيك وانطتني
شينطيني انطوا دأنطيك باعطي ينطيهم أعطيني أعطيت أعطاه سأعطيك لإعطاء
إعطاء أعطيها وانطاك وانطيها انطوهن انطيهن وأنطيهم أنطيها تنطيك أنطي
ودانطيه وانطيه وانطينا أنطيكم انطوه ينطيهن باعطيك أنطوني أعطانا بإعطاء
إنطيني ينطيكم وأعطيت أعطت وينطي وتنطيك شتنطيه شانطيك انطوكم انطونا
اعطينا اعطني ينطي ينطيح أعطوا اعطوني انطيها إنطيتهن أعطى انطتني
تنطيني انطيه انطيني انطيناك إنطيته انطيتني انطيتك انطي انطوها انطاني
وتنطيه انطيتها أعطي بأعطيك اعطونا إنطوا فانطاني

B Permission Letters



public full text

Report message · Block user

Tiba Zaki Abdulhameed Sep 28, 2018

Hi,
I am Tiba Zaki , an author of a paper that was published in the proceedings of
INTERNATIONAL CONFERENCE ON NATURAL
LANGUAGE, SIGNAL AND SPEECH PROCESSING
DECEMBER 05 and 06
Casablanca 2017
http://icnlssp.isga.ma/abouts-us/proceeding
While the proceedings are open access, I shared the paper full text publicly in research gate, so do I have the right? This is because I can't refer to the paper in my google scholar.
Thank you and best regards,
Tiba

Kamel Smaïli to you Sep 28, 2018

Hi Tiba

Yes you have the right to do it.

Wish all success in your work

Pr K. Smaïli

Tiba Zaki Abdulhameed Sep 28, 2018

Thank you so much. I appreciate your fast reply.
Wish you all the best,
Tiba

Figure B.1.: Permission to include "Tiba Zaki Abdulhameed, Imed Zitouni, Ikhlas Abdel-Qader, and Mohamed Abusharkh.Assessing the usability of modern standard Arabic data in enhancing the language model of limited size dialect conversations. Casablanca, Moroco, December 2017. International Conference on Natural Language, Signal and Speech Processing 2017."

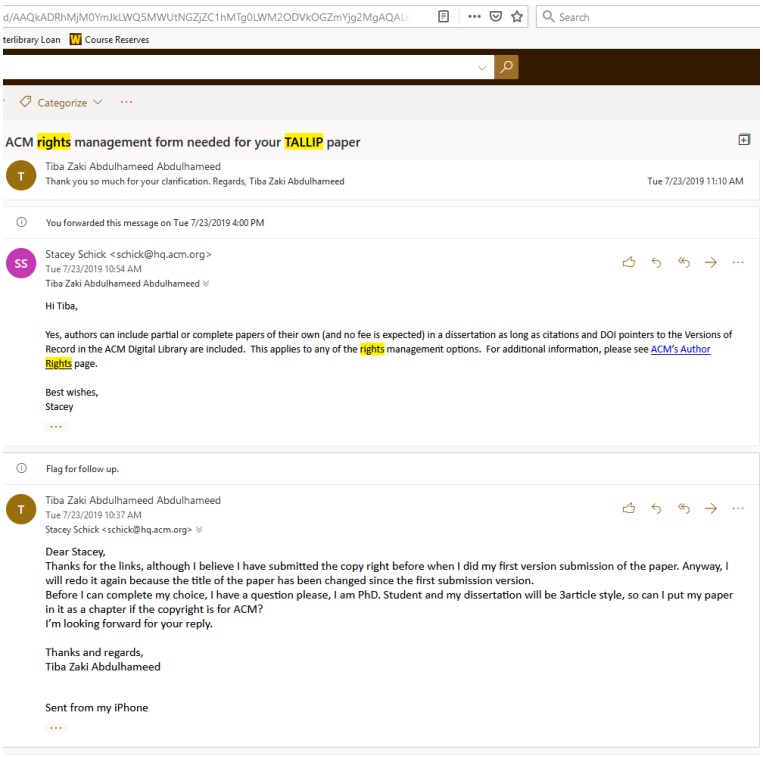


Figure B.2.: Permission to include "Tiba Zaki Abdulhameed, Imed Zitouni, and Ikhlas Abdel-Qader. "Wasf-Vec: Topology-based Word Embedding for Modern Standard Arabic and Iraqi Dialect Ontology." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 19, no. 2 (2019): 1-27."

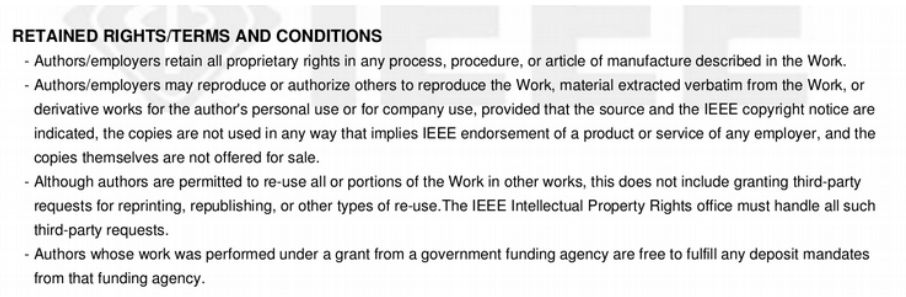


Figure B.3.: Permission to include "Tiba Zaki Abdulhameed, Imed Zitouni, and Ikhlas Abdel-Qader. Enhancement of the word2vec class-based language modeling by optimizing the features vector using pca. In 2018 IEEE International Conference on Electro/Information Technology (EIT), volume 2018-, pages 0866–0870. IEEE, 2018."