STEM-BASED ARABIC LANGUAGE MODELS EXPERIMENTS

Mohsen Moftah
Arab Academy for Science Technology and Maritime Transport
Heliopolis, Cairo
Egypt
Mohsen.moftah@barmagyat.com

Waleed Fakhr
Arab Academy for Science Technology and Maritime Transport
Heliopolis, Cairo
Egypt
waleedf@aast.edu

Sherif Abdou
Faculty Of Information Technology, Cairo University
Giza
Egypt
sherif.abdou@rdi-eg.com

Mohsen Rashwan
Faculty Of Engineering, Cairo University
Giza
Egypt
mrashwan@rdi-eg.com

Abstract

Arabic is one of the languages that are often described as morphologically complex. This nature of the Arabic language leads to rapid vocabulary growth which is accompanied by worse language model (LM) probability estimation and a higher out-of-vocabulary OOV rate. Morphology-based language models have been proposed to overcome such problems. In a morphology based language model the input text is analyzed and every word is split into a stem and affixes. In this paper, stem-based language models for Modern Standard Arabic (MSA) are developed and compared to the word-based model. The conventional word-based language model was considered as the baseline and stem-based language models are built and compared with the word-based one. For Stem-based language models, a number of manipulations were applied to the input data and new language models were built in each case and results were compared with both the baseline and the original stem-based language model.

Introduction

The classical approach in language model building is using the whole word. The input training text is used to build the model dictionary by extracting the unique words encountered within the input text. A standard statistical language model (LM) computes the probability of a word sequence $W = w_1, w_2, \ldots, w_T$ as a product of the conditional probabilities of each word w_i given its history, which is typically approximated by the one or two most recent words. Even with this limitation, the estimation of LM probabilities is challenging since many word contexts are observed infrequently or not at all. This is particularly problematic for morphologically rich languages, e.g. Turkish, Russian, or Arabic. Such languages have a high vocabulary growth rate, which results in high language model perplexity and a large number of out-of-vocabulary (OOV) words (Dimitra Vergyri et al, 2004**)**. A commonly used trigram model can be applied to a word based language model to achieve this goal. This approach fits well languages with a relatively strict word order and weak morphology. Inflective languages with relatively free word order require a more sophisticated approach to LM construction (Ilya Oparin and Andre Talanov 2007). In this paper we introduce different experiments in Arabic language modelling based on morphological analysis using Modern Standard Arabic (MSA).

Modern Standard Arabic Language

Modern Standard Arabic (MSA) is the literary standard across the Middle East and North Africa, and one of the official six languages of the United Nations. Most printed matter in the Arab World including most books, newspapers, magazines, official documents, and reading primers for small children is written in MSA.

Literary, Arabic is the official language of all Arab countries and is the only form of Arabic taught in schools at all stages. MSA can be considered a simplified version of the Classical Arabic in which speakers do not always observe the complicated rules of Classical Arabic. Both the structure and vocabulary were influenced by other languages. In writing, some authors try to use a style closer to the Classical Arabic others introduce new modern styles.

Arabic Morphology Analysis

Linguistically, there are many approaches of Morphology Analysis all of them have the same goal which is analyzing the structure of the word in a language. Hence, Morphology is concerned with the study of the internal structure of the word which is considered the smallest unit of syntax, a word is defined as: The unit of language that represents a concept which can be expressively communicated with meaning. In most, if not all,

languages words can be related to other words by rules. A speaker of a specific language recognizes these relations from tacit knowledge of the rules of word formation of that language. The rules understood by the speaker reflect specific patterns (or regularities) in the way words are formed from smaller units and how those smaller units interact in speech. In this way, another definition of morphology can be as the branch of linguistics that studies patterns of word formation within and across languages, and attempts to formulate rules that model the knowledge of the speakers of those languages. The smaller units that compose a word are called morphemes.

Based on the above definitions, a word consists of one or more morphemes which are linked more or less tightly together, and has a phonetic value. Morphemes may be stems, or affixes. The stem can be defined as: The part of word that is common to all its inflected variants.

Typically a word will consist of a stem and zero or more affixes, the affixes may be a *prefix* attached at the beginning of a stem or a *suffix* attached at the end of the stem. And can be expressed as

prefix-stem-suffix**

The stem is different from the *root* which is defined as*:* The primary lexical unit of a word which cannot be reduced into smaller constituents and does not appear on its own. Different words are derived from the same root. This also means that many stems may belong to the same root. Table (1) shows an example of a single root with many stems derived from it along with many words derived from the same stem.

The prefixes and suffixes could also be composite, for example the word "ليفاوضونهم" can be decomposed as shown in Table (2).

word	suffix	stem	prefix	root
أَكْتُبْ		كْتُب	أ	كتب
أَكْتُبُهُ	ـه	كْتُب	أ	كتب
أَكْتُبُهَا	ـهَا	كْتُب	أ	كتب
فَكَتَبَ		كَتَبَ	فَـ	كتب
الْكَاتِبِ		كَاتِب	الـ	كتب
كَكَاتِبْ		كَاتِب	كَـ	كتب
الْكَاتِبَاتِ	ـات	كَاتِب	الـ	كتب
الْكَاتِبَيْنِ	ـيْنِ	كَاتِب	الـ	كتب
اَلْكِتَابْ		كِتَاب	الـ	كتب
الْكِتَابَانِ	ـانِ	كِتَاب	الـ	كتب
الْكِتَابَهْ	ـة	كِتَاب	الـ	كتب
الْكُتَّابِ		كُتَّاب	الـ	كتب
الْكُتَّابْ		كُتَّاب	الـ	كتب
الْمَكْتَبْ		مَكْتَب	الـ	كتب
الْمَكْتَبَاتْ	ـات	مَكْتَب	الـ	كتب
الْمَكْتَبَهْ	ـة	مَكْتَب	الـ	كتب

Table 1: Example of words and stems for a single root

ليفاوضونهم				
prefix		stem	suffix	
لي		فاوض	ونهم	
ل	ي		ون	هم

Table 2: Example of prefix/suffix decomposition

Language Modelling

The goal of a language model is to determine the probability $\boldsymbol{P(w_l^n)}$ of a word sequence $\boldsymbol{w_l^n}$. This probability is decomposed as follows:

$$P(w_1^n) = \prod_{i=1}^{n} P(w_i / w_l^{i-1})$$

The most widely-used language models are n-gram models. In n-gram language models, we condition the probability of a word on the identity of the last (*n -1*) words. The choice of *n* is based on a trade-off between detail and reliability, and will be dependent on the available quantity of training data (Karima Meftouh et al 2008).

To establish the word n-gram language model, probability estimates are typically derived from frequencies of n-gram patterns in the training data. It is common that many possible word n-gram patterns would not appear in the actual data used for estimation, even if the size of the data is huge and the value of n is small. As a consequence, for rare or unseen events the likelihood estimates that are directly based on counts become problematic. This is often referred to as the data sparseness problem. Smoothing is used to address this problem and has been an important part in any language model (Xiaoyong Liu and W. Bruce Croft, 2004).

Language modeling is used in many natural language processing applications such as speech recognition, machine translation, part-of-speech tagging, parsing and information retrieval.

Evaluation of language models has typically been done using a measure called "perplexity. This measure is directly related to entropy. Entropy measures the average uncertainty present for a random variable. The more knowledge or structure a model captures, the lower the uncertainty, or entropy will be. Models with lower entropy can therefore be considered better. Mathematically, the perplexity *PP*, can be defined as the average number of possible words following any string of (*N-1*) words in a large corpus based on *N-gram* language model (L. Rabiner and B.H. Juang, 1993) *PP* is represented as follows:

$$PP = 2^{\frac{1}{n}\log_2(P(w))}$$

With *n* is the size of the test corpus (where W is the test data word sequence).

Data Preparation For Different Experiments

The data used in the experiments carried out are taken from Al-Ahram newspaper, a very popular Egyptian newspaper. The used text is 65K sentences containing 3.34 M words. The 65K sentences were divided into two parts, 45K sentences for models training and 25K sentences for testing. Both the training and test data were manipulated to fit a different experiment to build a language model. In our experiments we built five different Language Models as follows.

Word-based Language Model

This model will be the base line in comparing the results. In this model a 2,4,6-gram language models were built using the original sentences without any manipulation. The wordlist used as dictionary is the most frequently

used 5K words of the training data. An example of sentence is shown in Figure 1.

يمثل الدكتور سليمان حزين علامة مهمة في تاريخ التعليم في وطنه مصر ويصعب جدا أن تتكرر الفرصة التي اتيحت لسليمان حزين فهو من ناحية أول خريج في أول دفعة في أول كلية من أول جامعة حكومية حيث كان أول دفعته التي كانت أول دفعة تتخرج في جامعة فؤاد الأول بعدما درس أربع سنوات في هذه الكلية الجديدة من الجامعة الجديدة ومن الناحية الأخري فقد كان سليمان حزين المدير المؤسس لرابع جامعة مصرية وهي في ذات الوقت أول جامعة تمت نشأتها الحقيقية في عهد الثورة

Figure 1: Sample Sentence

Stem-based Language Model (Normal)

In this model the sentences are morphologically analyzed using RDI ArabMorpho®. The morphology analyzer decomposes the word and returns the prefix if exists, the stem, and the suffix if exists. The sentence is then reconstructed but each word is replaced by the pattern *prefix stem suffix*. A 2,4,6-gram language models were built using the most frequently used 5K stems of the training sentences as a dictionary. An example of reconstructed sentence is shown in Figure 2.

يـ مَثَّلَ الـ دُكْتُور سُلَيْمَان حُزَيِّن عَلَام ـة مُهِمّ ـة فِي تَارِيخ الـ تَعْلِيم فِي وَطَن ـة مِصْر وَيـ صْعُب جِدّ أَنْ تُـ تَكَرَّرَ الـ فُرْص ـة الـ لَتِي أُتِيحَ ـتْ لِـ سُلَيْمَان حَزِين فَـ هُوَ مَنْ نَاحِي ـة أَوَّل خِرِّيج فِي أَوَّل دَفْع ـة فِي أَوَّل كُلّ ـيَّة مِنْ أَوَّل جَامِع ـة حُكُوم ـيَّة حَيْثُ كَانَ أَوَّل دَفَعَ ـتْهُ الـ لَتِي كَانَ ـتْ أَوَّل دَفْع ـة تـ تَخَرَّجَ فِي جَامِع ـه فُؤَاد الـ أَوَّل بَعْدَمَا دَرَسَ أَرْبَع سَنَوَ ـات فِي هذِهِ الـ كُلّ ـيَّة الـ جَدِيد ـة مِنْ الـ جَامِع ـة الـ جَدِيد ـة وَ مِنْ الـ نَاحِي ـة الـ أُخْرَى فَـ قَدْ كَانَ سُلَيْمَان حُزَيِّن الـ مُدِير الـ مُؤَسِّس لِـ رَابِع جَامِع ـة مِصْر ـيَّة وَ هِيَ فِي ذَات الـ وَقْت أَوَّل جَامِع ـة تَمّ ـتْ نَشْء ـتهَا الـ حَقِيق ـيَّة فِي عَهِدَ الـ ثَوْر ـة

Figure 2: Sample Sentence for Stem-Based Model (normal)

Stem-based Language Model (reordered)

The same analysis done in the normal Stem-based Model is repeated. The sentence is then reconstructed but each word is replaced by the pattern *prefix suffix stem*. An example of reconstructed sentence is shown in Figure 3.

يـ مَثَّلَ الـ دُكْتُور سُلَيْمَان حُزَيِّن ـة عَلَام ـة مُهِمّ فِي تَارِيخ الـ تَعْلِيم فِي ـة وَطَن مِصْر وَيـ صْعُب جِدّ أَنْ تُـ تَكَرَّرَ الـ ـة فُرْص الـ لَتِي ـتْ أُتِيحَ لِـ سُلَيْمَان حَزِين فَـ هُوَ مَنْ ـة نَاحِي أَوَّل خِرِّيج فِي أَوَّل ـة دَفْع فِي أَوَّل ـيَّة كُلّ مِنْ أَوَّل ـة جَامِع ـيَّة حُكُوم حَيْثُ كَانَ أَوَّل ـتْهُ دَفَعَ الـ لَتِي ـتْ كَانَ أَوَّل ـة دَفْع تـ تَخَرَّجَ فِي ـه جَامِع فُؤَاد الـ أَوَّل بَعْدَمَا دَرَسَ أَرْبَع ـات سَنَوَ فِي هذِهِ الـ ـيَّة كُلّ الـ ـة جَدِيد مِنْ الـ ـة جَامِع الـ ـة جَدِيد وَ مِنْ الـ ـة نَاحِي الـ أُخْرَى فَـ قَدْ كَانَ سُلَيْمَان حُزَيِّن الـ مُدِير الـ مُؤَسِّس لِـ رَابِع ـة جَامِع ـيَّة مِصْر وَ هِيَ فِي ذَات الـ وَقْت أَوَّل ـة جَامِع ـتْ تَمّ ـتهَا نَشْء الـ ـيَّة حَقِيق فِي عَهِدَ الـ ـة ثَوْر

Figure 3: Sample Sentence for Stem-Based Model (reordered)

Stem-based Language Model (fixed normal)

The same analysis done in the normal Stem-based Model is repeated. The sentence is then reconstructed but each word is replaced by the pattern *prefix stem suffix*. In this model, if the word does not have a prefix the character "#" is inserted as a null prefix indicator, if the word does not have a suffix the character "&" is inserted as a null suffix indicator. In this way the pattern of the each word is fixed to *prefix stem suffix.* An example of reconstructed sentence is shown in Figure 4.

يـ مَثَّلَ & الـ دُكْتُور & # سُلَيْمَان & # حُزَيِّن & # عَلَام ـة # مُهِمّ ـة # فِي & # تَارِيخ & الـ تَعْلِيم & # فِي & # وَطَن ـة # مِصْر & وَيـ صْعُب & # جِدّ & # أَنْ & تُـ تَكَرَّرَ & الـ فُرْص ـة الـ لَتِي & # أُتِيحَ ـتْ لِـ سُلَيْمَان & # حَزِين & فَـ هُوَ & # مَنْ & # نَاجِي ـة # أَوَّل & # خِرِّيج & # فِي & # أَوَّل & # دَفْع ـة # فِي & # أَوَّل & # كُلّ ـيَّة # مِنْ & # أَوَّل & # جَامِع ـة # حُكُوم ـيَّة # حَيْثُ & # كَانَ & # أَوَّل & # دَفَعَ ـتْهُ الـ لَتِي & # كَانَ ـتْ # أَوَّل & # دَفْع ـة تـ تَخَرَّجَ & # فِي & # جَامِع ـه # فُؤَاد & الـ أَوَّل & # بَعْدَمَا & # دَرَسَ & # أَرْبَع & # سَنَوَ ـات # فِي & # هذِهِ & الـ كُلّ ـيَّة الـ جَدِيد ـة # مِنْ & الـ جَامِع ـة الـ جَدِيد ـة وَ مِنْ & الـ نَاجِي ـة الـ أُخْرَى & فَـ قَدْ & # كَانَ & # سُلَيْمَان & # حُزَيِّن & الـ مُدِير & الـ مُؤَسِّس & لِـ رَابِع & # جَامِع ـة # مِصْر ـيَّة وَ هِيَ & # فِي & # ذَات & الـ وَقْت & # أَوَّل & # جَامِع ـة # تَمّ ـتْ # نَشْء ـتهَا الـ حَقِيق ـيَّة # فِي & # عَهِدَ & الـ ثَوْر ـة

Figure 4: Sample Sentence for Stem-Based Model (fixed normal)

Stem-based Language Model (fixed reordered)

The same analysis done in the normal Stem-based Model is repeated. As in the previous model, the null prefixes and suffixes are replaced by '#" and "&" respectively. The sentence is then reconstructed but each word is replaced by the pattern *prefix suffix stem.* An example of reconstructed sentence is shown in Figure 5.

يـ & مثل الـ & دُكْتُور # & سُلَيْمَان # & حزن # ـة علم # ـة همم # & فِي # & أرخ الـ & علم # & فِي # ـة وطن # & مصر وَيـ & صعب # & جدد # & أَنْ تُـ & كرر الـ ـة فرص الـ & اَلَّذِي # ـتْ تيح لِـ & سُلَيْمَان # & حزن فَـ & هُوَ # & اَلَّذِي # ـة نحو # & أول # & خرج # & فِي # & أول # ـة دفع # & فِي # & أول # ـيَّة كُلّ # & مِنْ # & أول # ـة جمع # ـيَّة حكم # & حَيْثُ # & كون # & أول # ـتْهُ دفع الـ & اَلَّذِي # ـتْ كون # & أول # ـة دفع تـ & خرج # & فِي # ـه جمع # & فأد الـ & أول # & بعد # & درس # & ربع # ـات سنه # & فِي # & ذَا الـ ـيَّة كُلّ الـ ـة جدد # & مِنْ الـ ـة جمع الـ ـة جدد وَ & مِنْ الـ ـة نحو الـ & أخر فَـ & قَدْ # & كون # & سُلَيْمَان # & حزن الـ & دور الـ & أسس لِـ & ربع # ـة جمع # ـيَّة مصر وَ & هُوَ # & فِي # & ذُو الـ & وقت # & أول # ـة جمع # ـتْ تمم # ـتهَا نشأ الـ ـيَّة حقق # & فِي # & عهد الـ ـة ثور

Figure 5: Sample Sentence for Stem-Based Model (fixed reordered)

Language Models Building Experiments

To make a rational comparison among the built language models we fixed all the used parameters used for building. Table (3) shows the parameters used in our experiments.

Training Data	Test data	LM Order	Vocabulary Size
45 K sentences	20 K sentences	2,4,6 gram	5K word/stem

Table 3: Parameters used in building LMs

Another constraint applied to our experiments that we did not handle composite affixes (prefixes and suffixes).

Referring to Table 2, we consider the prefix as "لي" and the suffix as "ونهم" without further breakdown for prefixes and suffixes.

The comparison among the different models was based on calculating the model perplexity and the OOV rates. The Word-Based model was considered a starting point for which the perplexity and OOV rate were calculated and used as a reference. For Stem-Based models different manipulations were done for the input text, both training and test.

The first manipulation is reordering the decomposed word. The output of the morphology analyzer for every word is *prefix stem suffix* each of which is treated separately when building the language model. We swapped the position of the *stem* and the *suffix* to be *prefix suffix stem* the goal is when the model tries to predict a stem it takes its suffix into consideration. In other words, the suffix will be one of the *N-1* phonemes preceding the stem being predicted.

The second manipulation was fixing the *prefix stem suffix* pattern. Because it is not necessary that every word has a prefix and/or suffix this means that the distance between two consecutive stems will not be the same. The following example illustrates the idea" عَلَام ـة مُهِمّ ـة فِي تَارِيخ الـ تَعْلِيم" the stem "مُهِمّ " is preceded by the suffix "ـة " and followed by the prefix "ـة " of the next word while the word "تَارِيخ " is preceded by the stem "فِي ". To have a fixed pattern, null prefix and suffix characters are inserted. And the previous stream of words will be

"# عَلَام ـة # مُهِمّ ـة # فِي & # تَارِيخ & الـ تَعْلِيم &"

The same reordering manipulation was applied to fixed pattern text.

Experiments Results

Table (4) shows the results of the different experiments. For perplexity, we first compare the stem-based model with the word-based model as shown in the table and Figure (6) the perplexity dropped significantly in stem-based model with all orders of models. This also applies for OOV, where OOV dropped significantly in stem-based model compared with wor-based model.

For the perplexity of different stem-based manipulations:

1. For normal stem-based models, reordering in smaller orders of model had negative effect on perplexity (bigram) then it slightly improved perplexity in higher orders (hexagram).
2. Stem-based fixed models significantly improved perplexity compared with normal stem-based models for all orders.
3. For stem-based fixed models, reordering significantly improved perplexity in smaller orders models (bigram and fourgram), in higher orders hexagram improvement was insignificant.

For OOV rate of different stem-based manipulations Figure (8):

4. For normal stem-based models, reordering did not improve OOV rate.
5. Stem-based fixed models slightly improved OOV rate compared with normal stem-based models.
6. For stem-based fixed models, reordering insignificantly improved OOV rate.

Model	Perplexity			OOV
	bigram	fourgram	hexagram	
Word Based	423.6465	329.8535	314.8856	31.20%
Stem Based Normal	91.5471	55.8621	55.8035	7.06%
Stem Based Normal Reordered	93.8830	55.4447	54.3588	7.06%
Stem Based Fixed Normal	68.8679	32.9419	17.2794	4.70%
Stem Based Fixed Reordered	34.6543	27.3640	16.4796	4.27%

Table 4: Experiments Results

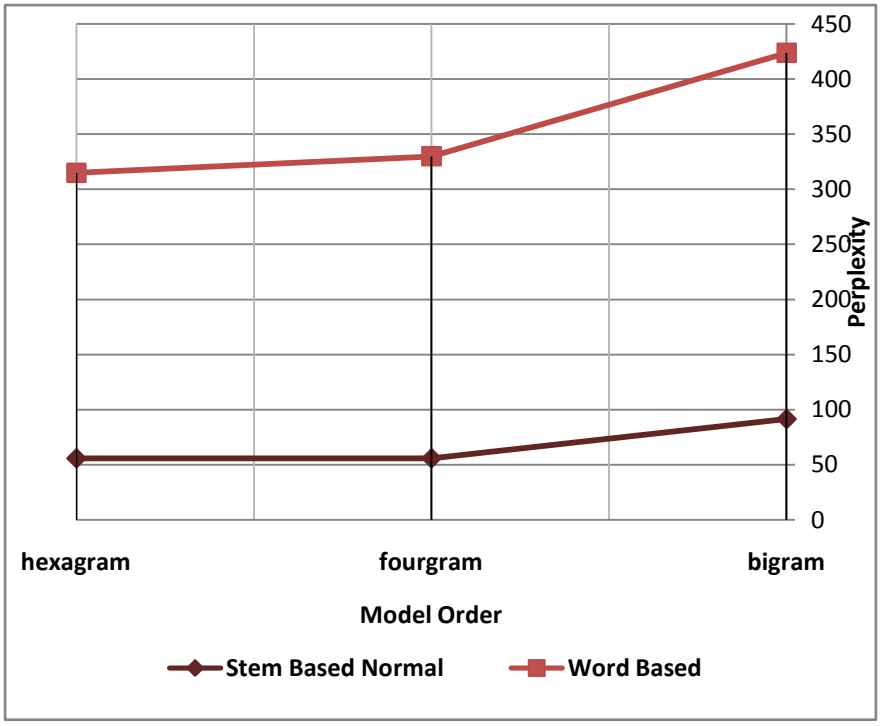

Figure 6: Perplexity of Word-Based and Stem-Based Models

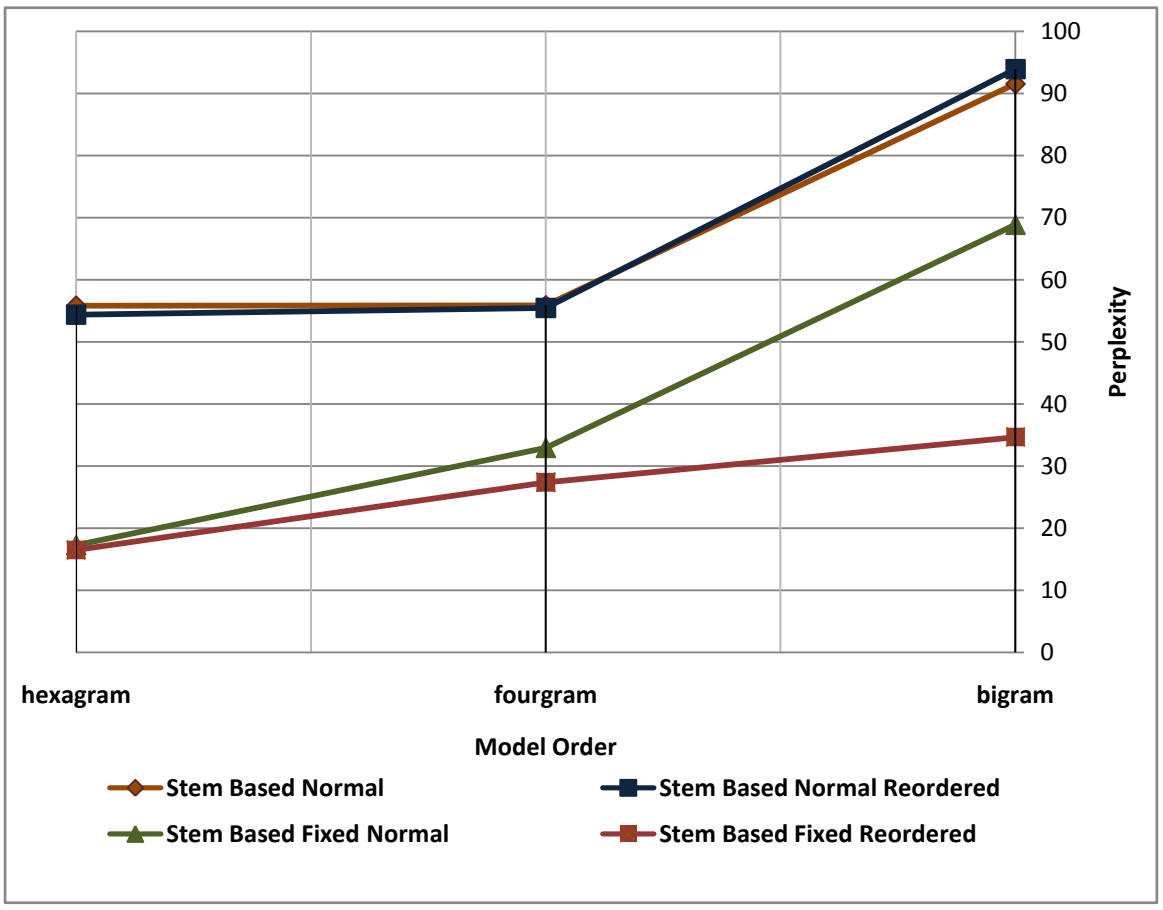

Figure 7: Perplexity of different stem-Based Models

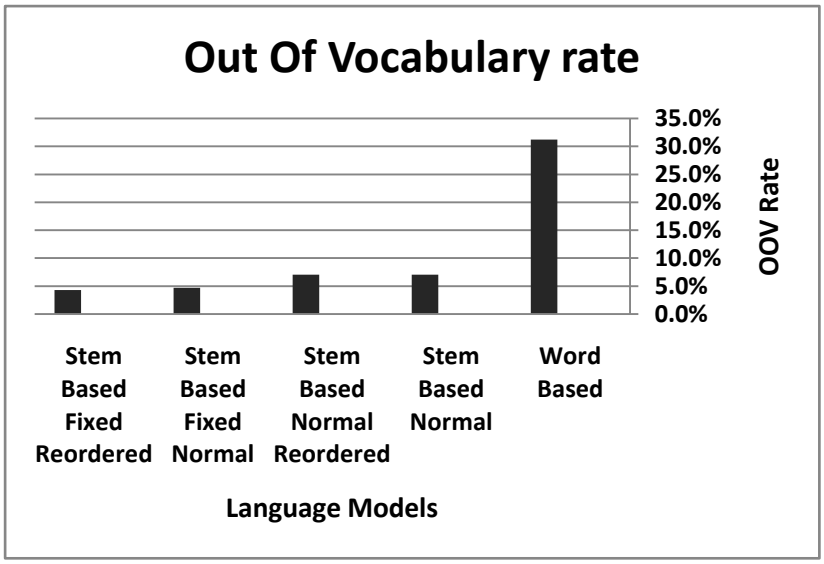

Figure 8: OOV Rate for Different Models

Conclusion and Future Work

In this paper different experiments were carried out for building Language Models for Arabic Language. The evaluation of the built language models was based on calculating perplexity and OOV rate. A basic conclusion is that using stem-based language models reduces the perplexity significantly than word-based models. For the different stem-based models, reordered model has a slight improvement over normal one.

Fixed stem-based models outperformed normal models in all cases. For fixed models, reordering has significantly improved the performance in low order n-grams, while it had slight improvement for higher orders.

Future work will include integrating these different models with an ASR engine and measuring their in- vivo performance. Also, investigating methods to combine word-based language models with stem-based language models either by interpolation or by a back off strategy.

Acknowledgements

A special thanks and appreciation to Research & Development International Company (RDI) for its support and for making both the data and the morphology analyzer (ArabMorpho) available to carry out this work.

Bibliographical References

Dimitra Vergyri (1), Katrin Kirchhoff (2), Kevin Duh (2), Andreas Stolcke (1) (2004). Morphology-Based Language Modeling for Arabic Speech Recognition
(1) SRI International, USA
(2) University of Washington, USA

Ilya Oparin and Andre Talanov (2007). "Stem-Based Approach to Pronunciation Vocabulary Construction and Language Modeling for Russian". Speech Technology Center. St. Petersburg, Russia.

Karima Meftouh, Kamel Smaili, Mohamed-Tayeb Laskri (2008). Arabic statistical language modelling. Badji Mokhtar University – Computer Science Department – BP 12 23000 Annaba – Algeria, INRIA-LORIA – Parole team – BP 101 54602 Villers Les Nancy – France

L. Rabiner and B.H. Juang (1993). Fundamental Of Speech Recognition (pp. 450). New Jersey: Prentice Hall.

Xiaoyong Liu and W. Bruce Croft. Statistical Language Modeling For Information Retrieval. Center for Intelligent Information Retrieval (2004). Department of Computer Science University of Massachusetts