MORPHAME-BASED ARABIC LANGUAGE MODELING FOR AUTOMATIC SPEECH RECOGNITION.

OLA M. ALI, HOWIDA A. SHEDEED & MOHAMED F. TOLBA
Scientific Computing Department
Ain Shams University
Cairo
EGYPT
olafern@yahoo.com, dr_howida@yahoo.com, fahmytolba@gmail.com

MAHMOUD GADALLA
Elect. Engineering Dept.
Military Technical Collage
Cairo
EGYPT
mahmoud_mtc@yahoo.com

Abstract: The main problems in Arabic speech recognition can be summarized as; Arabic phonetics, diacritization problem, grapheme-to-phoneme relation, and morphological complexity. These problems are considered challenges when designing an effective language model for Arabic language. This paper introduces a comparative investigation for morpheme-based language modeling techniques that tried to solve the problems of Arabic in Speech Recognition systems. The paper discusses the previous works that used both linguistic knowledge and statistical estimation in Morpheme-based language modeling as it pertains to Arabic and other morphologically rich languages. A proposed framework for Arabic language modeling using morphological analysis is also presented in this paper. Using our model, we hope to overcome the lack of generality of some previous systems, while simultaneously working to overcome the out-of-vocabulary (OOV) problem and addressing the problem of lacking resources.

Key-Words: Automatic speech recognition, language modeling, morpheme.

1 Introduction

The conversion of speech to text in automatic speech recognition presents a series of linguistic and computational challenges in any language. Some of these relate to language modeling (studying large amounts of text to learn about patterns of words in a language). [1] summarized the main problems in Arabic speech recognition, which include: Arabic phonetics, diacritization problem, grapheme-to-phoneme relation, and morphological complexity. To conclude, the three main language-specific challenges are:

1 -The morphological properties of Arabic such that, there are many more word forms in Arabic, related to each semantically coherent word, than in English. For example, the word forms: ‘استعمل’ (to use), ‘عامل’ (worker - factor), ‘عامل’ (to deal with), ‘عمل’ (laboratory), and more than 8 other forms are all looked up under the root word ‘عمل’. Each word form will occur less often than the more abstract word to which it is related.

2 -The regional dialects of Arabic differ in such ways that hinder the sharing of linguistic resources. Building NLP tools usually requires some amount of written text. Because most of Arabic dialects are primarily spoken and not written, acquiring resources for NLP tools is difficult. Table 1 lists some examples of the differences between Egyptian Colloquial Arabic (ECA) and Modern Standard Arabic.

<table>
<thead>
<tr>
<th>Change</th>
<th>MSA</th>
<th>ECA</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>/th/ → /s/</td>
<td>/tha/</td>
<td>/ta/</td>
<td>‘three’</td>
</tr>
<tr>
<td>inflections</td>
<td>/takllam/</td>
<td>/tikllam/</td>
<td>‘he speaks’</td>
</tr>
<tr>
<td>vocabulary</td>
<td>تايليا</td>
<td>تاربيئيزا</td>
<td>‘table’</td>
</tr>
<tr>
<td>word order</td>
<td>VSO</td>
<td>SVO</td>
<td></td>
</tr>
</tbody>
</table>

3- The orthography (written form) of Arabic lacks information about some vowel sounds. This is called the short-vowel problem. In order to build reliable acoustic models, the vowels are either given explicit symbols in the text, or the symbols representing consonants are modeled as consonant + vowel sequences. For example, the form "كتك" (ktb) can correspond to "كتاب" (kataba), or "كتب" (kutub), or 19 other forms in a large variety of contexts, which decreases predictability in the language model.

These problems lead to high (OOV) rates, and poor Language Model (LM) probabilities. A traditional approach to overcome these problems is to use a very large recognition vocabulary. Yet, still relatively high OOV rates are obtained. Moreover, the speech recognition system suffers from high resource requirements such as CPU time and memory. An alternative approach is to use morpheme-based LMs in order to lower the OOV
rate and perplexity, reduce data sparsity, decrease resource requirements and achieve lower Word Error Rates (WERs)[9].

The rest of the paper is organized as follows. Sec 2 presents a literature review for morpheme-based Arabic LMs. Sec 3 presents the comparative study for recent researches for morpheme based Arabic LMs. Sec 4 presents our proposed framework for Arabic LM. Conclusions are drawn in Sec 5.

2 Related work

Normally, morphemes are generated by applying morphological decomposition to words. In some cases morphological decomposition is based on linguistic knowledge as in [2], and in other cases it is based on unsupervised approaches like in [3]. Some of the linguistic methods make use of the Buckwalter Arabic Morphological Analyzer (BAMA) like in [4] and [5].

Recently, researches find that to overcome the data sparseness and to reduce the dependence of the traditional word-based LMs on the discourse domain, we can assign proper features (classes) to words and build LMs over those features. This yields better smoothing and, hopefully, better generalization to unseen word sequences. Possible approaches for incorporating word features into LMs are: stream-based LMs [6], class-based LMs [7] and factored LMs [8].

[9] Investigated the use of features derived for morphemes rather than word to combine the benefits of both morpheme level and feature rich modeling. The performance estimation of stream based, class-based and Factored LMs (FLMs) were compared over sequences of morphemes and their features. It was verified that, those feature-based LM techniques could be used in morpheme domain as efficient as in word domain. Morpheme-based LMs achieved better lexical coverage and reduced the problem of data scarcity. While the feature-based models tried to achieve better generalization to unseen word sequences.

[10] Struggled Automatic Speech Recognition (ASR) models to handle accented speech, particularly if the target accent was under-represented in the training data.

[11] developed ASR and statistical machine translation (SMT) systems. A number of different techniques were evaluated in the MT and SMT tracks.

3 Morpheme-based Language Modeling

Many of the world’s languages exhibit a degree of morphological complexity such as Turkish, Finnish, Estonian, German, and other languages in addition to Arabic. All of them use a wide variety of bound inflectional affixes, compound affixes, or compound words to create new terms. By counting each unique word form as a separate term in the lexicon, these morphological processes can produce a very large lexicon.

The goal of NLP is to provide some analysis for each term in a given text or utterance. When a term in a new text is unaccounted for a given training text, it is called an out-of-vocabulary (OOV) term. OOV terms typically do not receive an analysis, as no information is given for them in the model.

Table 2: Comparison across previous research of word accuracy improvements using morphemes, relative to comparable word models within the same study.

<table>
<thead>
<tr>
<th>Publication</th>
<th>Language</th>
<th>Morpheme/Model Type</th>
<th>Best Improvement (% Relative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afify et al. [21]</td>
<td>Iraqi</td>
<td>Affix</td>
<td>14.60</td>
</tr>
<tr>
<td>El-Desoky et al. [9]</td>
<td>MSA</td>
<td>MADA</td>
<td>12.64</td>
</tr>
<tr>
<td>Nguyen et al. [26]</td>
<td>MSA</td>
<td>Combination</td>
<td>11.70</td>
</tr>
<tr>
<td>Xiang et al. [12]</td>
<td>MSA</td>
<td>Affix</td>
<td>9.84</td>
</tr>
<tr>
<td>Choueiter et al. [25]</td>
<td>MSA</td>
<td>Affix, med vocab</td>
<td>7.57</td>
</tr>
<tr>
<td>Diehl et al. [19]</td>
<td>MSA</td>
<td>MADA</td>
<td>7.46</td>
</tr>
<tr>
<td>Sarikaya et al. [22]</td>
<td>Iraqi</td>
<td>Affix</td>
<td>7.45</td>
</tr>
<tr>
<td>Nguyen et al. [26]</td>
<td>MSA</td>
<td>Sakhr</td>
<td>6.36</td>
</tr>
<tr>
<td>Nguyen et al. [26]</td>
<td>MSA</td>
<td>Affix</td>
<td>3.72</td>
</tr>
<tr>
<td>Vergyri et al. [13]</td>
<td>Egyptian</td>
<td>FLM</td>
<td>3.34</td>
</tr>
<tr>
<td>Kirchhoff et al. [24]</td>
<td>Egyptian</td>
<td>FLM</td>
<td>3.11</td>
</tr>
<tr>
<td>El-Desoky et al. [9]</td>
<td>MSA</td>
<td>MADA, voweled</td>
<td>3.00</td>
</tr>
<tr>
<td>Kirchhoff et al. [24]</td>
<td>Egyptian</td>
<td>Combination</td>
<td>2.54</td>
</tr>
<tr>
<td>Kirchhoff et al. [24]</td>
<td>Egyptian</td>
<td>Stream, Class</td>
<td>1.94</td>
</tr>
<tr>
<td>Stolcke [29]</td>
<td>Levantine</td>
<td>Affix</td>
<td>1.41</td>
</tr>
<tr>
<td>Choueiter et al. [25]</td>
<td>MSA</td>
<td>Affix, large vocab</td>
<td>0.74</td>
</tr>
<tr>
<td>Emami et al. [23]</td>
<td>MSA</td>
<td>Rich NN</td>
<td>0.00</td>
</tr>
<tr>
<td>Creutz et al. [27]</td>
<td>Egyptian</td>
<td>Morfessor</td>
<td>-4.06</td>
</tr>
</tbody>
</table>

A successful way of approaching the OOV problem is to break the words into sub-word units (like morphemes) under the assumption that carefully chosen sub-word units have fewer unique...
types, thereby limiting the OOV problem [28]. In addition, well chosen sub-word units should repeat more often than whole words in a given text, making their frequency statistics more reliable. However, replacing words with sub-word units can result in new problems: shorter units often cause n-gram models to have a smaller amount of contextual information, and may be harder to model acoustically or semantically. To counteract these deficiencies, model designers may change parameters such as the order of n-gram model or the size of the sub-word unit.

Sub-word units often provide a good partial solution to the OOV problem in automatic speech recognition (ASR) ([9],[14], [2] ). Table 2 shows that there is a precedent for using morpheme based models. Many studies gain significant improvements in word accuracy by enhancing the models in this way. However, there are two studies that show no improvement, and for some studies where morphemes were helpful, the effect is only slight.

In [13], in building an Arabic ASR system, morphemes were discarded as the models grew large because “as more data became available the improvement from these models diminished”. Some publications indicate that the amount of benefit is highly dependent on the corpus. In [26], five evaluation sets were used; for the best model, the improvement in word accuracy across those five sets ranged from 2.2% to 0.9% absolute.

There does not seem to be a clear correlation between the complexity of the system or the size of the model and the utility of morphemes, according to the list in Table 2. Furthermore, there does not seem to be a correlation between complexity of morpheme derivation and success of the resulting model. For instance, [21] used the relatively simple affix definition method, also constraining predicted stems using the BAMA tool. It is highly successful within its domain of Iraqi Arabic. On the other hand, the FLM of [24], which take into account a multitude of linguistic data and sophisticated backoff algorithms, provided a much smaller improvement for Egyptian Arabic, as does the simple affix method when used with a large vocabulary on MSA data.

We might conclude that, each research in Table 1 proves the utility of morpheme models and morpheme derivation methods for that system. It is difficult if not impossible to predict whether a complex, simple, or lack of morpheme derivation method will be most useful for a particular implementation of Arabic ASR.

This section discusses the previous works that used both linguistic knowledge and statistical estimation to get morpheme derivation, as it pertains to Arabic.

3.1 Deriving Sub-Word Units using Linguistic Knowledge

The tools described in this section were designed to automatically produce morphological analyses of Arabic words. In all of these works, some computational structure, such as a finite state machine (FSM), encodes the information necessary for decomposing each word into morphemes or sub-word units that express a semantic or syntactic meaning.

[16] used linguistic information to create a morphological analyzer for Arabic. Labels from the Penn Arabic Treebank [17] were used to train multiple classifiers for morphological analysis. The treebank was created with the aid of the Buckwalter Arabic Morphological Analyzer (BAMA) [18], which used a dictionary of Arabic prefixes, stems, and suffixes, together with combination rules, to hypothesize all morphological analyses of a word. Human annotators marked the correct BAMA analysis for thousands of words in context in the treebank. The morphological analyzer in [16] was based on the multiple BAMA analyses and the correct choices given in the Penn Treebank, two rich sources of linguistic knowledge. These classifiers then provide the best analysis of each word. The resulting tool, called MADA, has recently been used successfully in several ASR systems for both morpheme generation and diacritization [9] and [19]. There was an inherent limitation with this tool in such a way that, it was built on Modern Standard Arabic (MSA) data and definitions.

[20] built an annotation tool that allowed annotators to manually choose the best morphological analysis of a given word. Each morpheme was also coded with a series of features describing its morphological and semantic properties.

3.2 Deriving Sub-Word Units using Statistical Information

The following research work makes more use of statistical estimation than linguistic knowledge to segment the morphemes that comprise each word. The goal of these methods is to produce morphemes useful to natural language processing tasks. However, heuristics are often used to ensure that the analyses do not stray too far from theoretical “correctness”.

[25] Used this method of morpheme derivation in ASR for MSA, and find that these morphemes are useful when the decoding vocabulary is small and
the OOV rate is large. Unsurprisingly, the effect is mitigated when there is a large decoding vocabulary that already counteracts the OOV problem. However, more recent studies including [26] have found that, the use of morpheme language models in Arabic ASR is useful even when the vocabulary is large.

[13] and [27] used only statistical information to derive sub-word units from words. Their model was language independent. To find sub-word units, the words in the lexicon were recursively split, and the cost of the model was re-calculated. Segmentations that lower the model cost were retained. It includes units common to many words are likely to be retained, as they tend to lower the cost of the model. No linguistic knowledge was taken into account.

3.3 Deriving Sub-Word Units using Linguistic Knowledge and Statistical Information

In the following studies, linguistic knowledge is combined with statistical estimation to derive sub-word units specifically for use in language modeling. In [25] Iraqi Arabic affixes were used to perform blind segmentation over the training corpus. Heuristics, including setting a minimum stem length and checking stem accuracy against the BAMA stem dictionary, limit the number of segmentations of each word. Linguistic knowledge of both Iraqi Arabic and MSA was thus applied. Each word in the training text is mapped to a single segmentation, and the resulting segments comprise the language model and pronunciation dictionaries. [25] found that, they were more successful at decoding with morphemes than decoding with words, as the morpheme vocabulary includes fewer OOV terms in the test set. The authors also found that, interpolating the morpheme model with a word model is successful; this was likely due to the increased contextual knowledge held by the word model. [26] used neural networks to include part of speech and vowel information as extra contextual data. Morphemes were derived using a three-part FSM that segments based on affixes, then refines the hypotheses based on dictionary entries and a segment based LM. The linguistic knowledge was a set of pre-defined affixes and a set of pre-defined stems. That work introduced a neural network language model to Arabic ASR, but it was found that the gain in word error rate is negligible.

[27] Described FLMs to be state-of-the-art in Arabic ASR. FLMs encode many types of information for each word, including morphological information where available. During the language model estimation and subsequent decoding, backoff for unseen n-grams occurs using any of the features encoded in the FLM. The lack of resources annotated corpora and tools - for dialectal Arabic make this method somewhat less viable. It may be for this reason that FLMs have not been highly visible in the Arabic NLP literature since the cited publication.

[28] used both linguistic knowledge and statistical estimation to derive morphemes from words. In that work the linguistic knowledge was the set of verbal patterns that combine with roots to form stems. In addition, a pre-defined set of affix characters included to limit the amount of over-generation. Frequencies of segment sequences were collected to statistically determine which are the most useful for further NLP processing. As in [17], finite state machines were used to encode the linguistic information. [28] determined the comparative usefulness of four morpheme derivation techniques for Arabic by building language models for ASR system with the resulting morphemes. Among four models and two dialects, it was found that, the utility of building morpheme-based language models was limited on that system. This conclusion does not negate previous literature showing that, word accuracy increases through the use of morpheme language models.

4 proposed framework for Arabic language modeling

In our proposed framework we intend to build a model for deriving morphemes from Arabic words using stem patterns, a feature of Arabic morphology. In this way, we hope to overcome the lack of generality of some previous systems, while simultaneously working to overcome the out-of-vocabulary problem and addressing the problem of lacking resources.

The structure of the ASR system is as shown in Figure 1.

The proposed model uses FLM with both bi-grams and tri-grams form morphemes. In FLM, words are viewed as vectors of K factors W = \( f^{1..K} \). The task then is to produce a statistical model over the resulting representation using a trigram approximation. The resulting probability model is as shown in equation (1):

\[
p(f_1^{1..K}, f_2^{1..K}, ..., f_T^{1..K}) \approx \prod_{t=1}^{T} p(f_t^{1..K} | f_{t-1}^{1..K}, f_{t-2}^{1..K})
\]  

A factor could be any feature of the word such as, morphological class, stem, root or even a semantic feature. The main idea of the model is to backoff to other factors when some word n-gram is not observed in the training data, thus improving the probability estimates.
For training data we intend to use King Abdulaziz City of Science and Technology (KACST) database for Arabic language sounds [30]. It is phonetically rich Arabic words contains a list of 663 phonetically rich words representing all Arabic phonemes, which are subject to all Arabic phonetic rules.

Figure 1: the structure of the ASR System

In the proposed model a SearchGraph is used to get the recognized Arabic word. As shown in Figure 2 SearchGraph is a directed graph composed of optionally emitting SearchStates and SearchStateArcs with transition probabilities. Each state in the graph can represent components from the Language Model, Dictionary or Acoustic Model.

Figure 2: Example SearchGraph of the words “sifr” and “wahid” which mean zero and one.

The used acoustic model should be built with word-based transcripts, using a vocalized pronunciation model if possible. To improve the base scores, one might look to improving the acoustic model by adjusting the short vowel predictions, triphone clustering, or adding more data. Given a working word-based language model, we propose to add morpheme sequences to the language model to increase the accuracy. It is then simple to use the one-to-one word-to-morphs mapping to transform the corpus into word and morph sequences before counting n-grams to build the language model.

If grammatical information is to be incorporated in other ways into the speech recognition system or output, then the work necessary to use more complex tools or the stem-based morpheme derivation method might be merited. If, on the other hand, a less informed model will suffice, the affix-definition method can be employed without great effort. [28] and [29] advised to use heuristics to limit the possible morpheme decompositions when using the affix method, in order to limit the number of spurious morpheme hypotheses. In addition, they advised to combine the morpheme hypotheses with the word hypotheses in some way. This can be done through interpolation of word-only and morpheme-only models, and through limiting decompositions to only the less-frequent words. So, we propose to use Frequent Words in the language model to limit the possible morpheme decomposition.

5 Conclusion
This paper introduces a comparative study for the morpheme-based language modeling techniques that solved the problems of Arabic in ASR systems. Also a proposed framework for morpheme-based Arabic LM is presented in this paper.

The problem of creating tools that work cross-dialectally remains a serious, important challenge, and the use of stem patterns to bridge the gap between dialects has not been ruled out as a possible fruitful path. Furthermore, the challenge of predicting short vowels in Modern Standard Arabic as well as colloquial Arabic remains an open question. We intend to solve these problems using our proposed model as mentioned in the previous section.

References


