INTERSPEECH 2009 BRIGHTON

# Investigating the Use of Morphological Decomposition and Diacritization for Improving Arabic LVCSR

*Amr El-Desoky, Christian Gollan, David Rybach, Ralf Schlüter, Hermann Ney*

Lehrstuhl für Informatik 6 – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{desoky,gollan,rybach,schluter,ney}@cs.rwth-aachen.de

## Abstract

One of the challenges related to large vocabulary Arabic speech recognition is the rich morphology nature of Arabic language which leads to both high out-of-vocabulary (OOV) rates and high language model (LM) perplexities. Another challenge is the absence of the short vowels (diacritics) from the Arabic written transcripts which causes a large difference between spoken and written language and thus a weaker connection between the acoustic and language models. In this work, we try to address these two important challenges by introducing both morphological decomposition and diacritization in Arabic language modeling. Finally, we are able to obtain about 3.7% relative reduction in word error rate (WER) with respect to a comparable non-diacritized full-words system running on our test set.
**Index Terms**: speech recognition, morphological decomposition, diacritization, Arabic

## 1. Introduction

Arabic is considered one of the most morphologically complex languages like Turkish, Korean, Russian, Finnish, Estonian and German [1, 2, 3]. In Arabic, words are derived from roots which may be three, four or in rare cases five letters long by applying templates to get stems and then attaching different affixes to obtain a high number of different surface forms. This huge lexical variety causes data sparsity problems and leads to high OOV rates and high LM perplexities. The traditional way to overcome this problem is to use a large recognition lexicon typically having several hundred thousands of full-words. However, we still have relatively high OOV rates compared to other languages. In addition to this, the ASR system suffers from high resource requirements such as CPU time and memory. For these reasons, the morphological decomposition of compound words into morphemes is proposed in order to lower the OOV rate and perplexity, reduce data sparsity, decrease the resource requirements and improve the final WER as well.

There are two main approaches to morphological decomposition, those approaches based on linguistic knowledge [4, 5, 6, 7], and those based on unsupervised methods [8, 9, 10, 11]. For Arabic language, some of the linguistic methods are based on Buckwalter Arabic Morphological Analyzer (BAMA) with some added constraints like in [5]. Some other methods start with a fixed set of affixes and decompose words into stems and affixes based on pattern matching also with some added constraints like in [5, 6]. In [6], an additional interpolation of non-decomposed and decomposed LMs is proposed to deal with dialectal Arabic. In [7, 12], a LM-based morpheme generator is used to perform decomposition, plus a morpheme lattice constrainer to reject illegal sequences of morphemes. On the other side, most of the unsupervised methods are based on the minimum description length principle (MDL) like in [11]. An unsupervised text segmentation software is available for this purpose called Morfessor 1.0 [8].

Arabic language is also a strongly consonantal language with only three vowels, each of which has a long and short form. Normally, Arabic text is written without vowelization which means that the short vowels and also the gemination and nunation marks (called diacritics) are not indicated in the written text [13]. Thus, for a given written word, there are several possible vowelizations that may happen during pronunciation. In traditional systems, the vowel information is not captured in the language models. Instead, a relatively high number of pronunciation variants are used during ASR search in order to fill the gap between the spoken and written language. Recently, it has been shown that modeling short vowels in Arabic LMs can improve performance even when producing the traditional non-vowelized output [13, 14]. But, this can exacerbate the problem of data sparsity. For this reason, combining morphological decomposition and diacritization may be useful.

It is shown in [13] that by using a very large vowelized vocabulary of more than 1.2 million words, and a LM with a vowelized component, the WER can be reduced significantly.

In our work, we investigate the use of morphological decomposition in Arabic LMs in combination with diacritization. Thus, we use LMs containing diacritized morphemes which, according to our knowledge, is a new approach. Finally, we could achieve good improvements in OOV rate, WER and LM perplexity over the traditional non-diacritized full-words systems.

The paper is organized as follows: In section 2, we describe our experimental setup. In section 3, we present our methodology of performing morphological decomposition and diacritization. Our experiments are discussed in section 4, while Section 5 gives conclusions.

## 2. Experimental setup

Our acoustic models are triphone models trained using 1100h of audio material taken from two domains: broadcast news (BN) and broadcast conversation (BC). The basic acoustic models are trained based on Maximum Likelihood (ML) method. Then, a discriminative training based on Minimum Phone Error (MPE) criterion is performed to enhance the models [15, 16].

Our language model training corpora consist of around 206 Million running full-words including data from Agile Arab text, FBIS, TDT4 and GALE BN and BC. In all our experiments, the vocabularies are selected out of the text corpora using the ML approach as described in [17], where the OOV rate is minimized over some part of the text data called ”held-out data”. The same

text corpora are used to estimate back-off N-gram LMs with modified kneser-Ney smoothing using SRILM toolkit [18].

Our speech recognizer works in 3 passes. In the first pass, within-word acoustic models are used with no adaptation. The second pass uses across-word models with Constrained Maximum Likelihood Linear Regression (CMLLR) adaptation. Then, a third pass with additional Maximum Likelihood Linear Regression (MLLR) adaptation is performed. In each of the three passes, a bigram LM is used to produce lattices then these lattices are rescored using a higher order LM (mostly 4-gram).

To evaluate the recognition performance, two small corpora from GALE data sets are selected as the dev and test sets for all our experiments. Each set consists of around 40 Minutes of audio data from 10 episodes by channels from: Abu Dhabi, Dubai, Iraq, Kuwait, Jordan and Syria, including BN and BC speech during January to March 2007.

# 3. Methodology

In this section we describe how decomposition and diacritization of text data are performed, followed by the affix sets and constraints examined in our work.

## 3.1. MADA tool

All our LM training data are firstly prepared using MADA 2.0 tool [19]. MADA is a morphological analyzer and disambiguator for Arabic. It is a tool capable of performing morphological tagging (disambiguation), diacritization [19, 20], and tokenization [21] of Modern Standard Arabic (MSA). As described in [19, 20], MADA 2.0 uses BAMA 2.0 to generate all possible analyses for words of a given sentence. Then, it applies classifiers for a complete set of morphological features to the words. Then, a combiner is used to rank potential word analyses returned by BAMA by using the output of the classifiers and then choose the highest ranked analysis. Finally, MADA is able to associate a complete morphological tag with words in context. These tags are used to produce robust diacritization and tokenization (decomposition) for the words of the sentence. In addition, MADA performs stem orthographic normalization, which chooses among alternate orthographic variants of words.

The MADA 2.0 code is slightly adapted in order to preprocess the LM training data to rewrite the words of the corpora in the following format:

$$W/P_1+ \quad P_2+ \quad ...P_n+ \quad ST \quad +S$$

where "$W$" is the full diacritized form of the word, "$P_1+$" up to "$P_n+$" forms an optional sequence of prefixes, "$ST$" is a mandatory stem of the word and "$+S$" is an optional suffix. The number of prefixes may range from 0 up to 3 prefixes per word and the number of suffixes is either 0 or 1 suffix per word while only the stem is a one mandatory part. The "/" character is simply a separator that separates the word from its decomposition. Both the full-word and the decomposition are diacritized. A typical example of this format is the following word written in Buckwalter transliteration (means "and inside it"):

$$wabidAxilihA/wa+ \quad bi+ \quad dAxili \quad +hA$$

In case that MADA fails to get a proper analysis for the word, the word is written between double at-marks "@@" with no decomposition or diacritization, such as: @@AljAy@@ which is a dialectal word means "the coming".

By using this format as the baseline format for the text corpora, we can easily define any data we want to use in vocabulary selection and LM training. For example, we can remove diacritical characters to obtain non-diacritized data, or we can properly concatenate prefixes to obtain one prefix per word or even we can back-off to the full-word form under certain conditions.

## 3.2. Affix sets

During the work of this paper, two different groups of affixes are examined which are:

- **Basic affixes:** (multi-prefixes are allowed)
  *Prefixes*: {Al, b, f, k, l, ll, w}.
  *Suffixes*: {h, hA, hm, hmA, hn, k, km, kmA, kn, nA}.
- **Compound affixes:** (only single prefixes are allowed)
  *Prefixes*: {Al, b, bAl, f, fAl, fb, fbAl, fk, fl, fll, k, kAl, l ll, w, wAl, wb, wbAl, wk, wkAl, wl, wll}.
  *Suffixes*: {h, hA, hm, hmA, hn, k, km, kmA, kn, nA}.

Both Affix sets have the same suffixes but different prefixes. An attached '+' sign to the end of prefixes and the start of suffixes is used to mark affixes in order to allow for easy recovery to the original words by attaching affixes with the corresponding stems. Additionally, an '@' sign is attached to prefixes that end with "Al" or "ll" to distinguish the prefixes which are followed by a solar consonant from others followed by a lunar consonant. This is because when a solar consonant comes after "Al" or "ll", the final "l" (called in this case solar "l") is not pronounced while the solar consonant is geminated. An example of solar "l" happens in the word "Al$ms" which means "the sun", while an example of lunar "l" is in the word "Alqmr" which means "the moon". Solar consonants are: {t, v, d, *, r, z, s, $, S, D, T, Z, l, n}, while the lunar consonants are: {>, <, b, j, H, x, E, g, f, q, k, m, h, w, y}.

## 3.3. Decomposition and diacritization constraints

While processing MADA output, some few constraints may be applied in order to rationalize the morphological decomposition and diacritization of words:

- 1. No decomposition is done for words with very short stems, typically less than or equal to 2 letters.
- 2. No decomposition is done for top N highly ranked decomposable full-words, where the ranks are assigned after applying a maximum likelihood based vocabulary selection procedure to the full-words data.
- 3. No diacritization is done for top M highly ranked words, where the ranks are assigned the same way as in constraint 2.

The first constraint is adopted to avoid very short stems wrongly obtained by MADA. While the efficiency of the second and third constraints is examined by varying the values of N and M throughout our experiments.

# 4. Experiments

Our recognition experiments are divided into two main groups. In group (1), we used a set of morphologically decomposed non-diacritized LMs. In group (2), we used LMs which are decomposed and diacritized at the same time.

The OOV rates recorded in our experiments are normalized as described in [4]. So that, the OOV rates are comparable regardless of the used morphological units. Thus,

$$OOV_{norm} = OOV * \frac{N_d}{N_{org}} \quad (1)$$

Where $N_d$ is the number of words in the decomposed data, and $N_{org}$ is the number of original words. In the following experiments, we comment only on the results of our test set. The running time and memory improvements are measured only for the third recognition pass.

### 4.1. Morphologically decomposed non-diacritized LMs

In Table 1, we summarize the results of the first group experiments. The first entry of the table is our baseline system with 256k non-diacritized full-words. The rest of the table presents a set of decomposition experiments using basic and compound affixes. The decomposed vocabulary size is fixed to 70k and the rescoring LM order is fixed to 4-gram. The number of decomposable full-words retained without decomposition (the value of N) is increased gradually starting from zero. The output decomposed hypothesis is re-joint into full-words before scoring.

Table 1: *Recognition results for morphologically decomposed non-diacritized LMs (BL: baseline; BA: basic affixes; CA: compound affixes; mrfs: morphemes; wrds: words).*

| Voc. | # mrfs | #full wrds | test 4-grm ppl | dev OOV/WER [%] | test OOV/WER [%] |
|---|---|---|---|---|---|
| BL | 0 | 256k | 344.3 | 1.08/11.8 | 1.33/13.6 |
| BA | 70k | 0 | 69.4 | 0.93/13.9 | 1.36/16.2 |
| CA | 70k | 0 | 75.2 | 0.93/13.9 | 1.36/16.0 |
| | 65k | 5k | 206.9 | 0.99/11.9 | 1.19/14.0 |
| | 60k | 10k | 243.4 | 1.03/11.9 | 1.19/14.0 |
| | 50k | 20k | 279.2 | 1.19/**11.6** | 1.33/**13.3** |
| | 40k | 30k | 293.6 | 1.46/11.7 | 1.66/13.6 |
| | 30k | 40k | 292.1 | 2.00/12.0 | 1.99/13.9 |

It can be seen that, the use of compound affixes during decomposition is more beneficial than using basic affixes. This was expected because the existence of multiple prefixes in LM training data rises the number of sequences of prefixes giving them high probability and thus leads to high insertion rates in the recognition output.

Although, the decomposed vocabulary size (70k) is smaller than the full-words vocabulary size (256k), still nice reductions can be seen in both OOV rate and WER. The system with 50k morphemes (stems and affixes) and 20k full-words gives the best reduction in WER which is 2.2% relative (0.3% absolute) compared to the baseline system, while the same OOV rate is retained (1.33%). In addition, a significant relative reduction in running time of about 36% is achieved (from 23.22xRT to 14.92xRT), beside a memory reduction of 53% relative.

Since morphemes and words are units of different lengths, then their optimal performance may occur at different n-gram orders [11]. For this reason, we add some extra recognition experiments recorded in Table 2 where higher order LMs (5 to 7-gram) are used for lattice rescoring instead of 4-gram. We use the best vocabulary we have from the previous set of experiments, with size of 70k (50k morphemes + 20k full-words) achieving an OOV rate of 1.33%.

We can see that the use of higher order LMs leads to less improvement. This is due to poor language model probability estimates as a result of the sparse data problem. For this reason, the 4-gram LM will be used for the rest of experiments.

Now, we perform additional two experiments where the decomposed vocabulary size is raised to 140k and 256k in order to better verify our improvement against the baseline system. The results are summarized in Table 3.

Table 2: *Recognition results for higher order rescoring LMs (vocabulary size = 70k with 20k decomposable full-words).*

| LM order | test LM ppl | dev (OOV=1.19) WER [%] | test (OOV=1.33) WER [%] |
|---|---|---|---|
| 5 | 277.7 | 11.5 | 13.5 |
| 6 | 277.9 | 11.5 | 13.4 |
| 7 | 277.8 | 11.5 | 13.5 |

Table 3: *Recognition results for larger vocabulary sizes.*

| # mrfs | #full wrds | test 4-grm ppl | dev OOV/WER [%] | test OOV/WER [%] |
|---|---|---|---|---|
| 120k | 20k | 288.8 | 0.62/11.5 | 0.88/13.2 |
| 236k | 20k | 297.4 | 0.43/**11.3** | 0.62/**13.1** |

It can be seen that the improvement in both OOV rate and WER persists for larger decomposed vocabulary sizes. The 256k decomposed vocabulary system achieves 3.7% relative reduction (0.5% absolute) in WER and 0.71% absolute reduction in OOV rate compared to the baseline system. Furthermore, a relative reduction of 13.6% is recorded for the LM perplexity beside a 23% relative reduction in recognition time.

### 4.2. Morphologically decomposed diacritized LMs

In Table 4, we summarize the results of the second group experiments where a set of partially decomposed vocabularies each of size 140k containing 120k morphemes and 20k decomposable full-words are used with partial diacritization. With partial diacritization we mean that the diacritization is excluded for the M top ranked words as stated in constraint 3 in Section 3.3. The number of non-diacritized words is increased gradually starting from 20k. The reason of starting from 20k is that the 140k vocabulary contains around 20k words for which MADA could not provide diacritized forms. Therefore, this is the minimum number of non-diacritized words we could have. All Affixes are kept non-diacritized while all their possible pronunciations are included in the lexicon. Multiple pronunciation variants are provided for each non-diacritized word or morpheme, while a single variant is provided for each diacritized one. This is the variant that corresponds exactly to the diacritized form. A normalized OOV rate is computed after removing diacritization while WER is recorded after re-joining affixes and removing diacritization from the output hypothesis.

We can see from Table 4 that even if we use a smaller (140k) decomposed and partially diacritized vocabulary, still nice improvement can be seen in WER compared to the baseline 256k full-words system in Table 1. Our best results are recorded with a vocabulary containing 120k non-diacritized entries beside 20k diacritized entries. We could achieve a relative WER reduction of approximately 3.0% (0.4% absolute) compared to the baseline system. We can also see that no improvement in WER could be achieved over the only decomposed vocabulary (restated at the end of Table 4) where the same WER is recorded (13.2%). Lastly, it is worth noting that the average number of pronunciations per word is lowered down from 3.9 in case of

Table 4: *Recognition results for morphologically decomposed diacritized LMs (number of decomposable full-words = 20k). Third column shows the number of effective non-diacritized words if diacritization is totally removed from the vocabulary.*

| #diac wrds | #non diac wrds | #eff. non-diac wrds | test 4-grm ppl. | dev OOV/WER [%] | test OOV/WER [%] |
|---|---|---|---|---|---|
| 120k | 20k | 85k | 313.3 | 1.03/17.9 | 1.17/19.7 |
| 100k | 40k | 88k | 277.9 | 1.01/13.7 | 1.15/15.6 |
| 80k | 60k | 94k | 285.2 | 0.96/12.4 | 1.11/14.7 |
| 40k | 100k | 112k | 280.1 | 0.80/11.9 | 0.99/13.4 |
| 20k | 120k | 125k | 282.3 | 0.72/11.8 | 0.95/**13.2** |
| 10k | 130k | 132k | 280.1 | 0.66/**11.7** | 0.93/13.3 |
| 5k | 135k | 136k | 280.1 | 0.62/11.8 | 0.91/13.3 |
| 0 | 140k | 140k | 288.8 | 0.62/11.5 | 0.88/13.2 |

non-diacritized lexicon to 3.5 in case of partially diacritized lexicon. This leads to 6.7% relative reduction in memory usage.

## 5. Conclusions

We have investigated the use of morphological decomposition, and a combination of decomposition and diacritization in Arabic LMs. The best results are achieved by using a morphologically decomposed non-diacritized vocabulary containing 20k full-words. By using additional 236k morphemes, a WER reduction of 3.7% relative (0.5% absolute) could be achieved, beside a significant reduction of about 23% in recognition time compared to a 256k baseline system of traditional full-words. Also, increasing the number of morphemes could decrease the WER correspondingly. We believe that the difficulties related to the use of diacritized LMs are due to the data sparsity problem. As a future work, we need more investigation for how to incorporate diacritization into LMs. One idea is to incorporate the morphological decomposition and the diacritization information into LMs using factored language models (FLM).

## 6. Acknowledgments

We thank the CADIM group at Columbia University for providing the MADA tool. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

## 7. References

[1] M. Adda-Decker, "A corpus-based decompounding algorithm for German lexical modeling in LVCSR," in *Proc. European Conf. on Speech Communication and Technology*, vol. 1, Geneva, Switzerland, Sep. 2003, pp. 257 – 260.

[2] K. Carki, P. Geutner, and T. Schultz, "Turkish LVCSR: towards better speech recognition for agglutinative languages," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Istanbul, Turkey, Jun. 2000, pp. 3688 – 3691.

[3] E. W. D. Whittaker and P. C. Woodland, "Particle-based language modelling," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 1, Beijing, China, Oct. 2000, pp. 170 – 173.

[4] B. Xiang, K. Nguyen, L. Nguyen, R. Schwartz, and J. Makhoul, "Morphological decomposition for Arabic broadcast news transcription," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Toulouse, France, May 2006, pp. 1089 – 1092.

[5] L. Lamel, A. Messaoudi, and J. Gauvain, "Investigating morphological decomposition for transcription of Arabic broadcast news and broadcast conversation data," in *Interspeech*, vol. 1, Brisbane, Australia, Sep. 2008, pp. 1429 – 1432.

[6] M. Afify, R. Sarikaya, H.-K. J. Kuo, L. Besacier, and Y. Gao, "On the use of morphological analysis for dialectal Arabic speech recognition," in *Interspeech*, vol. 1, Pittsburgh, PA, USA, Sep. 2006, pp. 277 – 280.

[7] G. Choueiter, D. Povey, S. Chen, and G. Zweig, "Morpheme-based language modeling for Arabic LVCSR," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Toulouse, France, May 2006, pp. 1053 – 1056.

[8] M. Creutz, "Induction of the morphology of natural language: Unsupervised morpheme segmentation with application to automatic speech recognition," Ph.D. dissertation, Helsinki University of Technology, Finland, 2006.

[9] J. Goldsmith, "Unsupervised learning of the morphology of a natural language," *Computational linguistics*, vol. 27, no. 2, pp. 153 – 198, Jun. 2001.

[10] Z. Harris, "From phoneme to morpheme," *Language, Linguistic Society of America*, vol. 31, no. 2, pp. 190 – 222, Jun. 1955.

[11] M. Creutz, T. Hirsimki, M. Kurimo, A. Puurula, J. Pylkknen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraclar, and A. Stolcke, "Morph-based speech recognition and modeling of out-of-vocabulary words across languages," *ACM Transactions on Speech and Language Processing*, vol. 5, no. 1, Dec. 2007.

[12] Y. Lee, K. Papineni, S. Roukos, O. Emam, and H. Hassan, "Language model based Arabic word segmentation," in *Proc. Annual Meeting of the Association for Computational Linguistics*, vol. 1, Sapporo, Japan, Jul. 2003, pp. 399 – 406.

[13] A. Messaoudi, J. Gauvain, and L. Lamel, "Arabic broadcast news transcription using a one million word vocalized vocabulary," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Toulouse, France, May 2006, pp. 1093 –1096.

[14] M. Afify, L. Nguyen, B. Xiang, S. Abdou, and J. Makhoul, "Recent progress in Arabic broadcast news transcription at BBN," in *Proc. European Conf. on Speech Communication and Technology*, vol. 1, Lisbon, Portugal, Sep. 2005, pp. 1637 – 1640.

[15] D. Rybach, S. Hahn, C. Gollan, R. Schlüter, and H. Ney, "Advances in Arabic broadcast news transcription at RWTH," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, vol. 1, Kyoto, Japan, Dec. 2007, pp. 449 – 454.

[16] D. Vergyri, A. Mandal, A. Stolcke, J. Zheng, M. Graciarena, D. Rybach, C. Gollan, R. Schlüter, K. Kirchhoff, A. Faria, and N. Morgan, "Development of the SRI/Nightingale Arabic ASR system," in *Interspeech*, vol. 1, Brisbane, Australia, Sep. 2008, pp. 1437 – 1440.

[17] A. Venkataraman and W. Wang, "Techniques for effective vocabulary selection in domain specific speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, vol. 1, Geneva, Switzerland, Sep. 2003, pp. 245 – 248.

[18] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 2, Denver, Colorado, USA, Sep. 2002, pp. 901 – 904.

[19] N. Habash and O. Rambow, "Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop," in *Proc. Annual Meeting of the Association for Computational Linguistics*, vol. 1, University of Michigan, USA, Jun. 2005, pp. 573 – 580.

[20] N. Habash and O. Rambow, "Arabic diacritization through full morphological tagging," in *Proc. Human Language Technology Conf. of the North American Chapter of the ACL*, vol. Companion, Rochester, NY, USA, Apr. 2007, pp. 53 – 56.

[21] N. Habash and F. Sadat, "Arabic preprocessing schemes for statistical machine translation," in *Proc. Human Language Technology Conf. of the North American Chapter of the ACL*, vol. 1, New York, USA, Jun. 2006, pp. 49 – 52.