IMPROVED MORPHOLOGICAL DECOMPOSITION FOR ARABIC BROADCAST NEWS TRANSCRIPTION

Tim Ng, Kham Nguyen†, Rabih Zbib⋆, and Long Nguyen

BBN Technologies, 10 Moulton Street, Cambridge, MA 02138, USA
† Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA
⋆ Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA
{tng,knguyen,rzbib,ln}@bbn.com

ABSTRACT

In this paper, we show the progress for Arabic speech recognition by incorporating contextual information into the process of morphological decomposition. The new approach achieves lower out-of-vocabulary and word error rates when compared to our previous work, in which the morphological decomposition relies on word-level information only. We also describe how the vocalization procedure is improved to produce pronunciations for some dialect Arabic words. By using the new approach, we reduced the word error by 0.8% absolute (4.7% relative) when compared to the baseline approach.

Index Terms— Speech recognition, Arabic, morphological decomposition

1. INTRODUCTION

Arabic is a morphologically rich language. A very large recognition lexicon would be needed in order to achieve a reasonable out-of-vocabulary (OOV) rate for a large-vocabulary Arabic Speech-to-Text (STT) system. Morphological decomposition of Arabic has been used in factored language models ([1], [2] and [3]) in order to reduce the OOV rate and alleviate the issue of training data sparsity in Arabic STT systems. A small improvement of around 2% to 3% relative has been reported. In our previous work ([4]), we compared different morphological decomposition algorithms based on word-level statistics for both acoustic and language modeling. The best decomposition algorithm achieved a significant reduction in OOV rate, and ultimately an 8.7% relative reduction in WER.

Although a significant improvement in WER was shown as a result of using morphological decomposition, the algorithms considered only word-level information. Contextual information, however, should play a role in morphological decomposition. Simple pattern matching is usually not enough to determine whether a string of characters is actually an affix or is part of the word stem. For example, the word "Alty"[1] could mean either "which" or "my machine" depending on the context. In the former case, it should not be decomposed, while in the latter, it should be decomposed into the stem "Alt" ("machine") and the suffix "y" ("my"). In this paper, we further improve the morphological decomposition in [4] by taking the contextual information into account. We show a further reduction in OOV rate and an improvement in recognition results. The work was part of the work for the GALE Phase 3 Evaluation.

Short vowels are usually not written in Arabic text. The LDC Buckwalter morphological analyzer is commonly used to vowelize the words for a pronunciation-based recognition dictionary. The analyzer is designed to process Modern Standard Arabic (MSA) words, although dialect words occur commonly too. Missing pronunciations of frequent dialect words prevents them from being included in the acoustic model training and the recognition dictionary of a pronunciation-based STT system. In this paper, we also describe improvements to the Buckwalter analyzer that allow it to handle a portion of dialect words.

The paper is organized as follows. In Section 2, we describe the training and test data used. We introduce an improved vocalization procedure in Section 3. The new morphological decomposition approach is presented in Section 4, and in Sections 5 and 6, we present the recognition system and experimental results. The paper concludes in Section 7.

2. TRAINING AND TEST DATA

The acoustic training data used in this paper consists of 1433 hours of audio data. It includes 43 hours of data from the FBIS corpus, 67 hours from TDT4, 50 hours of Iraqi Arabic data, and 1273 hours from GALE Phases 1, 2 and 3 Releases. The language models are trained on a corpus of 1.72 billion-word text. The data, which is shared among the GALE

This work was supported under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022

[1] Arabic examples are in Buckwalter format.

community, includes data from the Gigaword Arabic corpus and data downloaded from the Web by Cambridge University, LIMSI, BBN and Sakhr.

Five test sets are used in this work. The test sets shown in Table 1 are the un-sequestered GALE 2007 evaluation data, Eval07, which includes 67 episodes aired in Dec 2006 with a total duration of 3.04 hours; the development sets designed by LDC for the 2007 and 2008 GALE Evaluation, called as Dev07 and Dev08; and at6 and ad6 which were constructed as tunning and development sets for the GALE 2006 evaluation.

Test set	Epoch	Episodes	Duration (hours)
Eval07	Dec 2006	65	2.84
Dev07	Nov 2006	55	2.60
Dev08	May 2007	67	3.04
at6	Nov 2005, Jan 2006	23	6.48
ad6	Nov 2005, Jan 2006	22	5.94

Table 1. Details for the five test sets

3. IMPROVED VOCALIZATION

Arabic text is usually written without short vowels, so a certain written word can be pronounced in more than one way. One common method to produce pronunciations for Arabic words is by vowelizing them using the LDC Buckwalter morphological analyzer. The Buckwatler Analyzer, however, is developed for MSA words. It does not deal with dialect Arabic words, which appear frequently in our train and test data, especially in the Broadcast Conversation genre. To address the problem of missing pronunciations for dialect words, we incorporate dialect affixes into the Buckwalter Analyzer, allowing pronunciations for a portion of the dialect words to be generated. Table 2 shows the dialect affixes used to extend the Buckwalter Analyzer. Pronunciations from the manually vocalized corpora (The Arabic TreeBank corpora and the LDC Iraqi Lexicon) are also included.

A master dictionary of 1.2 million words is built from the union of pronunciations for words vowelized by the improved Buckwalter Analyzer and the pronunciations found in the manually vowelized data. A normalization procedure similar to the one described in [4] is applied to each word. The average number of pronunciations in the dictionary is around 4 per word. The phoneme set consists of 36 phonemes.

4. MORPHOLOGICAL DECOMPOSITION USING CONTEXTUAL INFORMATION

As we mentioned before, previous work on using morphological decomposition is based on word-level information. In this paper, we use an Arabic morphological analyzer that uses context information to determine the part of speech of each word in context. This information is then used to determine which word is decomposable. A given word is at most decomposed into one prefix, one stem and one suffix.

Prefixes	Suffixes
H Hn Ht Hy b bA bhAl bn bt btn by byn hAl t wbA wbhAl wbn wbt wbtn wby wbyn wtn wyn yn	Athm Atm Atn h hA hm hn k lA lh lhA lk lkm lkn lm ln lnA ly m n nA w wA whA wk wky wkm wkn wnA wny

Table 2. The dialect affixes that are incorporated into the Buckwalter Analyzer

It was shown in [4] that keeping the most frequent decomposable words unchanged benefits the recognition performance. Hence, a list of the 128K most frequent decomposable words, that we call the "blacklist" is constructed. The decomposable words are decomposed as follows:

- If the word is in the blacklist, keep unchanged.
- Else if no prefix, decompose into stem, suffix.
- Else if no suffix, decompose into prefix, stem.
- Else if (prefix + stem) is in the blacklist, decompose into (prefix + stem) and suffix.
- Else if (stem + suffix) is in the blacklist, decompose into prefix and (stem + suffix).
- Else decompose into prefix, stem and suffix.

4.1. Affixes

Table 3 lists the set of affixes used in the new approach, and the set affixes used in the baseline ([4]). There are 12 prefixes and 34 suffixes in the baseline system, while 24 prefixes and 13 suffixes are used in the contextual system. Also, to model the phonological rule where "Al" is assimilated into the following letter when it's a "sun" letter (e.g. "AlTAlb" *the student* is pronounced "ATTAlib"), we create variants of the prefix "Al" that correspond to each "sun" letter.

4.2. Pronunciation for Stems and Affixes

The pronunciations for stems are looked up in the master dictionary, while the pronunciation for the affixes were created manually. There are certain stems that are found in the acoustic model training data but whose pronunciations cannot be found in the master dictionary. In order to fully utilize all of the available acoustic model training, we obtain pronunciations for the stems through some automatic vowelization.

Approach	Prefixes	Suffixes
Baseline	Al, bAl, fAl, kAl, ll, wAl, b, f, k, l, s, w	An, h, hA, hm, hmA, hn, k, km, kn, nA, ny, t th, thA, thm, thmA, thn, tk, tkm, tm, tnA, tny, tynA, wA, wh, whA, whm, wk, wkm, wn, wnA, wny, y, yn
Contextual	Al, w, b, l, wAl, f, s, k, wb, wl, ws, fAl, wbAl, wll, kAl, fl, fs, wk, fb, fbAl, fll, wkAl	h, hA, hm, nA, y, hmA, ny, k, km, hn, kn, ky, kmA

Table 3. The affi xes used in the baseline and contextual morpheme systems

4.3. OOV Reduction

In [4], we reported a significant reduction in OOV rate as a result of using morphological decomposition. In this work, we further reduce the OOV rate using the new decomposition algorithm. To measure the OOV rate reduction, we pick 300K most frequent words from the LM data. We construct the blacklist from the top 128K words, and decompose the remaining words using both decomposition approaches.

The OOV rates of the baseline decomposition approach and that of the new one are shown in Table 4. The OOV rates for both systems are normalized according to the following equation:

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

where N_d is the number of words in the decomposed data and N_{org} is the number of original words. The ratio between these two is called the decomposition rate. As Table 4 shows, by using the contextual information in the decomposition process, we reduce the OOV rate from 0.94% to 0.60% on Dev08. The OOV rate reduction reflects the usefulness of the context information for morphological decomposition.

Decomp	Decomp Lex	OOV rate
Baseline	265K	0.94
Contextual	258K	0.60

Table 4. OOV rate for Dev08 against two decomposition approaches based on a 300K most frequent word lexicon (Decomp: Decomposition; Lex: Lexicon)

5. RECOGNITION SYSTEM

The recognition process in our system consists of three stages: speaker clustering, feature extraction and decoding.

5.1. Speaker clustering and Feature extraction

Voice activity detection is not performed since manual audio segmentation is given for the GALE Phase 3 evaluation. Online speaker clustering ([5]) and feature extraction are performed on the manually-marked audio segments. The length of each speech frame is 25 ms, with a frame rate of 100 frame/sec. For each frame, 14 perceptual linear prediction (PLP) [6] derived cepstral coefficients and energy are extracted. The Long Span Features (LSF) [7] are computed. The 9 successive frames of steady features (centered at the current frame) are concatenated. This block of features is projected onto a 60-dimensional feature space using Linear Discriminant Analysis (LDA).

5.2. Decoding

The decoding strategy of the recognition system used here is similar to that described in [4] and [8]. Three decoding stages, one unadapted and two adapted are used in the system.

The multi-pass search approach is applied in each decoding stage. The unlikely hypotheses are pruned in the forward pass in which a State Tied Mixture (STM) acoustic model and a bigram language model are used. The backward pass with a state clustered within-word quinphone acoustic model and a trigram language model is then performed on the space shrunk by the forward pass. The lattices that are output from the backward pass are rescored using a state-clustered cross-word quinphone model. Finally, the nbest lists from the lattice rescoring are reordered using an unpruned 4-gram LM. Speaker adaptation using LSF is described in [9]. The acoustic models in this paper are discriminatively trained using Minimum Phone Frame Error (MPFE) [10].

The recognition lexicon is derived from all the morphemes (decomposed words) in the acoustic training transcripts as well as those that occur at least 30 times in the language model training corpus. We end up with a lexicon of 284K morphemes.

The language model training corpus is partitioned into 25 groups according to their genre and sources. A language model is estimated on each of the 25 groups using the modified Kneser-Ney smoothing. The 25 language models are then linearly combined with weights optimized on a union set of at6 and Dev08 using the EM algorithm.

6. EXPERIMENTAL RESULTS

It was shown that the number of unique words that cannot be vocalized in at6 and ad6 can be reduced by 33% using the new vocalization procedure. To further investigate the effect, we trained two Maximum Likelihood (ML) phoneme systems using the new and old pronunciation dictionaries.

It is shown in Table 5 that, by using the new vocalization procedure with the same 333K vocabularies, 0.2% absolute reduction in WER is observed for Dev08.

System	Dev08
Old Vocalization	13.7
New Vocalization	13.5

Table 5. WER for the ML phoneme systems using old and new pronunciation dictionaries

The MPFE recognition results for the phoneme and the morpheme systems using the two different decomposition approaches are shown in Table 6. The vocalization procedure described in Section 3 is used for all of the systems. The 390K decoding vocabularies used in the phoneme system is from the word list as described in Section 5.2, and the 289K-morpheme recognition lexicon in the baseline morpheme system is also derived from the same word lists using the approach described in [4] with a 128K blacklist. The contextual system is the one described above. The system configuration and training are similar for the three systems.

The two morpheme systems are significantly better than the phoneme system for most of the test sets. In comparison to the baseline morpheme system, a 0.2% absolute increase in WER is observed for Dev08 using the contextual system, but 0.1% to 0.8% absolute improvement is obtained for the other test sets. Since Dev08 was used as a part of the heldout set in the language model training and as a tunning set for the recognition system, the performance on the other test sets is more relevant to us.

System	at6	ad6	Dev07	Eval07	Dev08
Phoneme	18.8	16.9	10.6	11.6	12.1
Baseline	18.1	17.1	10.3	11.1	11.6
Contextual	17.6	16.3	10.2	10.8	11.8

Table 6. WER for the phoneme and the two different morpheme systems

To get a better understanding of the effect of the new morphological decomposition approach, we looked at the hypotheses in which the contextual morpheme system performs better. A typical such example is shown in Table 7.

In the example, the words "fyHv" (encourages) and "llAyrAnyyn" (to the Iranians) were mis-recognized by the baseline morpheme system, while they are recognized correctly in the new morpheme system as a result of decomposing them into prefixes and stems.

7. CONCLUSION

We have introduced a method to incorporate contextual information into morphological decomposition. By using the new morphological decomposition, the reduction in WER can be as significant as 0.8% absolute (4.7% relative) when compared to our baseline morpheme system.

Reference:	AmA AlAmm AlmtHdp fyHv AmynhA AlEAm ElY AltfAwD wbnAG Alvqp wysdy AlnSyHp llAyrAnyyn
English:	but the_nations the_united encourages its_secretary general on the_negotiation and_building the_trust and_he_gives the_advice to_the_Iranians
Translation:	As for the UN, its Secretary General encourages negotiations and building the confi dence and he gives advice to the Iranians
Baseline:	AmA AlAmm AlmtHdp f_ **yxS** AmynhA AlEAm ElY AltfAwD wbnAG Alvqp wysd AlnSyHp **l AlAyrAnyyn**
Contextual:	AmA AlAmm AlmtHdp f_ yHv AmynhA AlEAm ElY AltfAwD wbnAG Alvqp wysd AlnSyHp ll_ AyrAnyyn

Table 7. Example for the hypotheses in which the contextual morpheme system is better off

8. REFERENCES

[1] D. Vergyri et al, "Morphology-based language modeling for arabic speech recognition," *Proc. ICSLP*, 2004.

[2] D. Vergyri et al, "Development of a conversational telephone speech recognition for levantine arabic," *Proc. Eurospeech*, pp. 1613–1616, 2005.

[3] K. Kirchhoff et al, "Novel approaches to arabic speech recognition: report from the 2002 johns-hopkins summer workshop," *Proc. ICASSP*, pp. 344–347, 2003.

[4] B. Xiang and K. Nguyen et al, "Morphological decomposition for Arabic broadcast news transcription," *ICASSP*, 2006.

[5] D. Liu and F. Kubala, "Online speaker clustering," *ICASSP*, May 2004.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87(4), pp. 1738–1752, 1990.

[7] B. Zhang, S. Matsoukas, and R. Schwartz, "Long span features and minimum phoneme error heteroscedastic linear discriminant analysis," *Proceedings of EARS RT-04 Workshop*, 2004.

[8] M. Afify and L. Nguyen et al, "Recent progress in Arabic broadcast news transcription at bbn," *Eurospeech*, 2005.

[9] T. Ng and B. Zhang et al, "Progress in the BBN 2007 Mandarin speech to text system," *ICASSP*, 2008.

[10] J. Zheng and A. Stolke, "Improved discriminative training using phone lattices," *InterSpeech*, September 2005.