Development of the SRI/Nightingale Arabic ASR system

D. Vergyri ¹, A. Mandal ¹, W. Wang ¹, A. Stolcke ¹, J. Zheng ¹, M. Graciarena ¹, D. Rybach ², C. Gollan ², R. Schlüter ², K. Kirchhoff ³, A. Faria ⁴, N. Morgan ⁴

¹ SRI International
² RWTH Aachen University
³ University of Washington
⁴ International Computer Science Institute
dverg@speech.sri.com

Abstract

We describe the large vocabulary automatic speech recognition system developed for Modern Standard Arabic by the SRI/Nightingale team, and used for the 2007 GALE evaluation as part of the speech translation system. We show how system performance is affected by different development choices, ranging from text processing and lexicon to decoding system architecture design. Word error rate results are reported on broadcast news and conversational data from the GALE development and evaluation test sets.

Index Terms: speech recognition, large vocabulary, Arabic

1. Introduction

The goal of the Global Autonomous Language Exploitation (GALE) program is to develop computer software techniques to analyze, interpret, and distill information from speech and text in Arabic and Chinese. In order to achieve the high performance targets set by the program, it has been necessary to drastically improve the accuracy of the first stage of GALE systems, which involve automatic speech recognition (ASR) in the source language. In this paper we focus on the Arabic ASR system work. Since the data of interest was from broadcast news (BN) and broadcast conversations (BC) we concentrated on Modern Standard Arabic (MSA), the most common Arabic dialect used in public broadcasts.

The morphological complexity of Arabic and other language features, like the lack of short vowels in standard orthography, introduce new requirements for the design of the ASR systems. The characteristics of the Arabic language are described in detail in [1]. Previous work that dealt with the problems of MSA ASR can be found in [2], [3], [4].

This paper describes the following aspects of system design, and reports their impact on the system’s performance: data preprocessing, lexicon and pronunciation probability generation, training data size, vocabulary size, sentence segmentation, front end features and discriminative training for acoustic models, and system architecture including system combination and cross adaptation. Our system includes components from two recognition engines: Decipher™ from SRI and the speech recognition system from RWTH Aachen. Results are reported for various GALE datasets that include both BN and BC speech.

2. Data

For acoustic training we used the 16 kHz Arabic data distributed by the Linguistic Data Consortium (LDC) with quick transcriptions: 45 hours of FBIS, 85 hours of TDT-4, 580 hours of GALE BC, and 740 hours of GALE BN data. We experimented with unsupervised training on an additional 3000 hours of untranscribed data, but these experiments showed ultimately no gains and are not included here. We also excluded all data sampled at 8 kHz that were distributed by LDC, since experiments showed no benefit from its use.

3. Text Processing and Lexicon generation

3.1. Text Data Processing

All text data was processed using the MADA¹ tool from Columbia University [6], which uses a statistical morphological tagger to fix common errors or ambiguities in the Arabic orthography, like the presence of Hamzas above or below the letters Alef, Yeh and Waw, and the marking of final Yeh and Teh Marbuta. MADA also provides full word diacritization, which was used as a means for developing the pronunciation lexicon. Other post-processing scripts were used to expand numbers and dates in the text data to words, remove nonlexical tags and punctuation marks, and normalize the orthography of some frequent names.

We did not compare our data processing method with that of other GALE sites (e.g., some sites do full normalization of Alefs and final Yeh), but we found that improved text processing had a significant effect on the accuracy of our models. Table 1 compares the word error rate (WER) results using MADA_1 (scripts released in 2005) versus MADA_2 (released in 2007 as v1.8). MADA_1 was based on the Buckwalter Arabic Morphological Analyzer² (BAMA) v.1 while MADA_2 used BAMA v.2, and the tagger was trained on more training data. The acoustic models used for this

<table>
<thead>
<tr>
<th>Text Processing</th>
<th>Eval06</th>
<th>Dev07</th>
</tr>
</thead>
<tbody>
<tr>
<td>MADA_1</td>
<td>36.3</td>
<td>25.7</td>
</tr>
<tr>
<td>MADA_2</td>
<td>35.7</td>
<td>25.2</td>
</tr>
</tbody>
</table>

Table 1. Effect of improved text processing on WER.

² LDC distributions LDC2002L49 and LDC2004L02
transcripts, of about 300K words. The final LM used a 600K balanced held-out mixture of GALE BN and BC manual about 1.1B words. Interpolation weights were tuned on a CMU web data, CU web data, and UN data, for a total of (manual transcriptions and web data), Arabic Gigaword, sources were FBIS, TDT-4, GALE BN and BC transcripts which were then interpolated for the final LM. The training specific LMs, each with modified Kneser-Ney smoothing, undiacritized MADA-processed data. We trained source-

4.1. Language Model Training

The LM was trained using the SRILM toolkit [8] with the undiacritized MADA-processed data. We trained source-specific LMs, each with modified Kneser-Ney smoothing, which were then interpolated for the final LM. The training sources were FBIS, TDT-4, GALE BN and BC transcripts (manual transcriptions and web data), Arabic Gigaword, CMU web data, CU web data, and UN data, for a total of about 1.1B words. Interpolation weights were tuned on a balanced held-out mixture of GALE BN and BC manual transcripts, of about 300K words. The final LM used a 600K vocabulary and was entropy-pruned to 40M bigrams for lattice generation decoding. For rescoring purposes a much bigger 4-gram was used with about 100M bigrams, 73M trigrams, and 73M four-grams. These LMs were used in both the SRI and the RWTH systems.

We also experimented with morphological-class LMs and factored LMs, but as more data became available the improvement from these models diminished, and they were not used in the final system.

4.2. Vocabulary Selection

To deal with the high morphological variability of Arabic, we had to enlarge the vocabulary far beyond what is typically used for English ASR. We used the approach described in [9] to choose the vocabulary, using the same held-out set as for LM tuning. Table 3 shows the effect of vocabulary size on OOV (out-of-vocabulary) rate and WER, using all the available data for LM.

4.3. Effect of LM Training Data Size

Using about half the available LM data (about 500M words from Gigaword, UN and the first 2 years of GALE) we found that the performance of the LM (after 4-gram rescoring) degrades about 1-1.5% absolute on various test sets. The experiment was performed using a 500K vocabulary.

5. Acoustic Modeling

5.1. SRI Acoustic Models

As described in Section 6.3, the SRI system uses multiple models at different steps: non-cross-word (non-cw) models are used for first-pass lattice generation, and for lattice rescoring. Cross-word (cw) models are used only for lattice rescoring and N-best list generation.

5.1.1. Front Ends

We used an MFCC front end with 16 kHz sampling rate and a PLP one with 8 kHz sampling rate, to help decorrelate errors for system combination. (The 8-kHz front end is motivated by the fact that some of the BC data is narrowband and/or noisy.) The MFCCs were augmented with multilayer perceptron (MLP) features as described in [10]. The non-cw MFCC and the PLP front ends have 13 cepstral coefficients while the cw MFCC has 19. They all use 1st, 2nd and 3rd order derivatives, with HLDA for reduction to 39-dimensional feature vectors. Table 4 shows the performance of the various models on different test sets. The PLP model without MLP features does not perform as well as the MFCC models, but because of the 8-kHz front end it performs well on the BC part of the data.
5.1.2. Discriminative Training

All models were trained with discriminative minimum phone error training (MPFE) [11]. We can see the gains over ML training in Table 4. We also experimented with using feature-level minimum phone error (fMPE) training [12],[13]. We found that the MLP features, which are already discriminatively trained, do not combine well with fMPE (there was practically no improvement from fMPE). That was the reason we decided not to include MLP features for the PLP front end and to use fMPE instead. We see that the PLP+fMPE models are comparable to the MFCC+MLP non-cw models after discriminative training. The cross-word MFCC19+MLP models performed the best.

5.1.3. Gender Dependent Models

Since we did not have data annotated for speaker gender, we used a gender prediction GMM trained on English data, to predict speaker gender in both training and test data. Then we performed MPE-MAP adaptation on the final discriminatively trained models, using the single-gender data only. This gave about 0.3% absolute improvement for individual models.

5.1.4. Duration Models

We used the approach described in [14] to train duration models for each acoustic model in our system. Single multivariate Gaussians are used to model phone duration features in words and triphones, normalized for a speaker’s average speaking rate. We employ these duration models for rescoring all N-best lists used in system combination (see Section 6.3). On average, duration modeling contributed a relative gain of about 0.2% absolute.

5.2. RWTH Acoustic Models

The RWTH system consists of two subsystems using two different acoustic models that share the same acoustic front end. MFCCs (16 cepstral coefficients) normalized by cepstral mean and variance normalization are augmented with a voicing feature. An LDA matrix projects the concatenation of 9 consecutive feature vectors in a sliding window to 45 components. As in the SRI MFCC models, this reduced feature vector is augmented with phoneme posterior features estimated by a neural network [15]. The two acoustic models used differ in their treatment of cross-word context: one subsystem uses within-word models only. In training, speaker variations are compensated by applying VTLN to the MFCC filterbank and speaker adaptive training (SAT) based on CMLLR. The models were enhanced by performing discriminative training with the MPE criterion.

5.3. RWTH ASR Architecture

Speaker-independent within-word models are used in the initial recognition pass. The output of this system is used to estimate the text-dependent CMLLR transforms for both subsystems. Segment clusters obtained by generalized likelihood ratio clustering with a BIC-based stopping criterion act as speaker labels for the speaker adaptation methods applied. The second recognition pass is carried out using the CMLLR transformed features and acoustic models trained with SAT. The lattices produced by the two systems are rescored using the full 4-gram language model, while a pruned bigram LM is used for generating the lattices. The acoustic models of both systems are adapted to this rescored output using CMLLRR on the means. We used a cross-adaptation scheme to benefit from having two subsystems. The results of the final recognition pass are again rescored using the larger LM. Eventually, the lattices produced are combined by a method based on min-fWER decoding [15]. Intermediate results for both systems are given in Table 5.

6. System Architecture

6.1. Acoustic Segmentation

Acoustic segmentation was applied in tasks where no manual transcription was available. Our segmentation method operates on the output of a speech recognizer. Several segment features are used to optimize the complete segmentation of a recording. A detailed description of this method can be found in [16]. In experiments where we used the non-cw system without LM rescoring and cross adaptation, we achieved a WER of 16.1% on Dev07 using the described segmentation. This result is quite close to the 15.9% achieved when using manual segmentations. However, there is still potential for improvement, which becomes apparent when splitting the manual segmentation before and after overlapping speech, yielding an error rate of 15.4%.

6.2. RWTH ASR Architecture

Figure 1 illustrates the design of the SRI ASR system. We generated lattices with the non-cw model twice (before and after adaptation) and cross-adapted all models used for final lattice rescoring to outputs of the other systems. The best system (MFCC19+MLP) was cross-adapted to the rover-combination of the other two systems. The results for different stages of the system are given in Table 6.
and system optimization for MT purposes.

system, factored LMs, partial diacritized vocabulary for LM modeling, unvowelized acoustic models as an alternate system includes improved pronunciation generation and difference of less than 1 point of BLEU or TER scores in all results obtained using the output of our system to those performance, we compared the machine translation (MT) Since the goal of the GALE project was speech-translation and system combination results. Our system obtained a WER different enough features to achieve good cross-adaptation with the needs of the new language. The different models had lexicon development and increased vocabulary size to deal with our English ASR experience, but we paid special attention to vocabulary ASR system for MSA. The system was based on We have described the development process of a large phone-loop adaptation Hypos for output or adaptation Generate or rescoring lattices

Figure 1. SRI System Architecture

Table 6. WER results at different system stages. Rover3 includes SRI-only systems A, C and D, while Rover5 includes also the two RWTH systems E and F.

<table>
<thead>
<tr>
<th>Output</th>
<th>Eval06 (manual seg.)</th>
<th>Dev07 (manual seg.)</th>
<th>Eval07 (automatic seg.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BN</td>
<td>BC</td>
<td>all</td>
</tr>
<tr>
<td>A</td>
<td>20.7</td>
<td>33.1</td>
<td>26.7</td>
</tr>
<tr>
<td>B</td>
<td>19.3</td>
<td>31.1</td>
<td>25.2</td>
</tr>
<tr>
<td>C</td>
<td>19.1</td>
<td>29.9</td>
<td>24.4</td>
</tr>
<tr>
<td>D</td>
<td>19.8</td>
<td>31.1</td>
<td>25.4</td>
</tr>
<tr>
<td>E-cw</td>
<td>20.4</td>
<td>31.8</td>
<td>26.0</td>
</tr>
<tr>
<td>F-cw</td>
<td>18.8</td>
<td>29.0</td>
<td>23.8</td>
</tr>
<tr>
<td>Rover3</td>
<td>18.2</td>
<td>27.9</td>
<td>23.0</td>
</tr>
</tbody>
</table>

Table 7. The final system WER on Eval07. Manual segments were available only for the BN part.

8. Acknowledgments

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

9. References


