NOVEL APPROACHES TO ARABIC SPEECH RECOGNITION: REPORT FROM THE 2002 JOHNS-HOPKINS SUMMER WORKSHOP

Katrin Kirchhoff[1], *Jeff Bilmes*[1], *Sourin Das*[2], *Nicolae Duta*[3], *Melissa Egan*[4], *Gang Ji*[1], *Feng He*[5], *John Henderson*[6], *Daben Liu*[3], *Mohamed Noamany*[3], *Pat Schone*[7], *Richard Schwartz*[3], *Dimitra Vergyri*[8]

[1]University of Washington, [2]Johns Hopkins University, [3]BBN, [4]Pomona College, [5]Swarthmore College, [6]MITRE, [7]Department of Defense, [8]SRI

ABSTRACT

Although Arabic is currently one of the most widely spoken languages in the world, there has been relatively little speech recognition research on Arabic compared to other languages. Moreover, most previous work has concentrated on the recognition of formal rather than dialectal Arabic. This paper reports on our project at the 2002 Johns Hopkins Summer Workshop, which focused on the recognition of dialectal Arabic. Three problems were addressed: (a) the lack of short vowels and other pronunciation information in Arabic texts; (b) the morphological complexity of Arabic; and (c) the discrepancies between dialectal and formal Arabic. We present novel approaches to automatic vowel restoration, morphology-based language modeling and the integration of out-of-corpus language model data, and report significant word error rate improvements on the LDC Arabic CallHome task.

1. INTRODUCTION: PROBLEMS IN ARABIC ASR

Arabic is currently the sixth most widely spoken language in the world with an estimated number of 250 million speakers. Despite this fact there has been little research on Arabic speech recognition compared to other languages of similar importance (e.g. Spanish or Mandarin). Most previous work on Arabic ASR has concentrated on developing recognizers for Modern Standard Arabic (MSA), which is a formal linguistic standard used throughout the Arabic-speaking world and is employed in the media (e.g. broadcast news), lectures, courtrooms, etc. Current recognizers for Arabic broadcast news are capable of achieving word error rates of 15-20% [1, 2]. However, MSA is only one among many varieties of Arabic - most informal, everyday communication is carried out in one of the regional dialects, of which there are four main types: Egyptian, Levantine, North African and Gulf Arabic. Initial work on developing ASR systems for dialectal Arabic was performed within the framework of the 1996/97 NIST benchmark evaluations on the CallHome task for different languages, including Egyptian Colloquial Arabic (ECA). The best performance obtained in those evaluations was 61% word error rate (WER) [3]. More recent systems obtain around 56% WER on the same task, which is still significantly higher than the word error rates on CallHome data in other languages.

Our work at the 2002 Johns Hopkins Summer Research Workshop focused on the recognition of conversational, dialectal speech as exemplified in the CallHome task. We addressed three prominent problems in Arabic ASR:

Script Representation: The Arabic alphabet only contains letters for long vowels and consonants. Short vowels and other pronunciation phenomena, like consonant doubling, can be indicated by diacritics (short strokes placed above or below the preceding consonant). However, Arabic texts are almost never fully diacritized and are thus potentially unsuitable for recognizer training. First, accurate acoustic model training is difficult when the identity and location of short vowels is unknown. Second, the absence of this information leads to many identical-looking word forms (e.g. the form كتب (ktb) can correspond to *kataba*, *kutub*, or 19 other forms) in a large variety of contexts, which decreases predictability in the language model.

Morphological Complexity: Arabic has a rich and productive morphology which leads to a large number of potential wordforms. This increases the out-of-vocabulary rate and prevents the robust estimation of language model probabilities.

Dialectal vs. Formal Speech: Arabic dialects are primarily oral languages; written material is almost invariably in MSA. Therefore, there is a serious lack of language model training material for dialectal speech.

In the following section we will describe the task and baseline systems, followed by descriptions of our approaches to automatic vowel restoration (Section 3), factored language modeling (Section 4) and using out-of-corpus text data (Section 5).

2. CORPUS AND BASELINE SYSTEMS

We used the only standardized corpus of dialectal Arabic currently available, the LDC CallHome (CH) corpus of Egyptian Colloquial Arabic. This is a collection of phone conversations between native speakers of ECA (mostly family members and friends). The corpus is divided into 80, 20, and 20 conversations for training, development and evaluation, respectively, corresponding to approximately 14, 3.5 and 1.5 hours. The corpus is accompanied by transcriptions in two formats: standard Arabic script without diacritics and a "romanized" version, which is close to a phonemic transcription. In past NIST evaluations the romanized versions were used as the official reference standard; however, there is growing consent among Arabic native speakers and speech system developers that producing and evaluating script recognition output would be more adequate. Romanized Arabic is unnatural and difficult to read for native speakers; moreover, script-based recognizers (where acoustic models are trained on graphemes rather than phonemes) have performed well on Arabic ASR tasks in the past [1, 2].

In order to assess the potential loss in performance when ignoring the additional phonetic information present in the romanized transcriptions we used two baseline systems for our work, one trained on script and another trained on the romanized transcriptions. Both systems were developed at BBN and are modifications of the Oasis broadcast news recognizer [2]. Differences to the original system include the use of the hand-segmentations provided in the LDC transcriptions rather than automatic segmentations; cepstral mean subtraction based on entire conversation sides rather than individual utterances; vocal tract length normalization, and LPC smoothing. Initial training was performed on the 80 conversation training set. The vocabulary size for both system was close to 14K; trigram language model perplexities on the development set were

AlHmd_llh	kwIsB	w	AntI	AzIk
llHamdulillA	kuwayyisaB	wi	inti	izzayik

Fig. 1. Example alignment of transliterated script (top row) and romanized word forms (bottom row).

Status in training set	% in test set	% error in test	% of total error
unambiguous	68.0	1.8	6.2
ambiguous	15.5	13.9	10.8
unseen	16.5	99.8	83.0
total	100.0	19.9	100.0

Table 1. Analysis of errors for baseline automatic romanizer.

508 (script) vs. 449 (romanized).
The script-based system obtained a WER of 59.9% on the 1996 evaluation set whereas the romanized system obtained 55.8% (evaluated against the script and romanized transcriptions, respectively). We thus see that ignoring the phonetic information available in the romanized transcriptions significantly degrades performance. The addition of another 20 training conversations released by LDC just before the workshop reduced the WER slightly in both systems, to 59.0% and 55.1%, respectively.

3. AUTOMATIC ROMANIZATION

The above results indicate that it would be advantageous to have large amounts of romanized training data for the development of future Arabic ASR systems. Since producing romanized transcriptions is costly and error-prone it would be preferable to romanize large amounts of undiacritized script data automatically. Some automatic diacritization tools are commercially available but have been developed for MSA; no tools are available for dialectal speech. Experiments with a commercial diacritizer for MSA showed a high performance on MSA texts (85%-95% accuracy) but poor results on CallHome data (53% accuracy). We therefore investigated the possibility of bootstrapping a statistical romanization tool based on a small amount of aligned script and romanized word forms (an alignment example is shown in Figure 1, in order to apply the resulting tool to a larger amount of script data.

For out initial experiments we used 40 conversations (the "eval97" and "h5_new" sets) for training the predictor and the 80 ASR training conversations for testing. As a simple baseline method we used maximum-likelihood unigram prediction, i.e. hypothesizing the romanized form most often seen in the aligned script/romanized training data. 80.1% of all script forms were romanized correctly by this method; an analysis of the baseline results is shown in Table 1. A large percentage of script forms (68%) have only a single romanized equivalent in the training set. Almost all of these are predicted correctly (the remaining errors are caused by words having only a single form in the training set but multiple forms in the test set). 86.1% of all ambiguous forms in the training set are predicted correctly. Not surprisingly, most errors are due to unknown script forms that were not seen in the training set; developing strategies for romanizing unseen words thus is crucial for good performance. To create possible romanizations for an unknown script form S_1, we find the closest corresponding script form, S_2, in the training set (based on a weighted Levenshtein distance measure), record the sequence of edit operations necessary to convert S_2 to its corresponding romanized form, R_2, and apply the same sequence of operations to S_1 to create a romanized form, R_1, for it. This procedure is inspired by the fact that many words of the same morphological class in Arabic share short vowel patterns. An example derivation is shown in Table 2. Further improvements to this method were obtained by using the top n (n=40) instead of the single best match for S_1, and by reversing the most frequent confusion patterns found from a statistical analysis of the romanization output on a held-out development set. Romanization accuracy increased from 80.1% to 83.5% overall and rose from 0.2% to 20.6% when measured just on the unknown script forms. For speech recognizer training, even partially correct romanized forms may be helpful as they can still contribute to better acoustic model training. We therefore measured character prediction accuracy in addition to word accuracy and found that it increased from 58.6% to 76.9%.

Step	Description	Example
1	S_1, unknown script form	tbqwA
2	S_2, closest known script form	ybqwA
3	R_2, corresponding rom. form	yibqu
4	edit operations for $S_2 \rightarrow R_2$	ciccrd
5	R_1, new rom. form	tibqu

Table 2. Example of romanization of unseen script form. Edit operations are copy, insert, replace, and delete.

Training	Eval reference	
	romanized	script
script	N/A	59.0
autoromanization	58.5	57.5
true romanization	55.1	54.9

Table 3. WER(%) obtained by training on script, true romanization and automatic romanization, evaluated against. Note that output from a script-trained recognizer cannot sensibly compared against romanized references (the WER would be 92.9%).

For our final speech recognition experiment we used the output of the automatic romanizer on the 100 training conversations used for our best system (see above). In this experiment there was an overlap of 20 conversations between romanization training and test set. Since novel word forms were now present in the training transcriptions due to incorrectly romanized words, pronunciation models had to be created for them. Although this was fairly straightforward due to an almost one-to-one correspondence between Arabic phonemes and graphemes used in the romanized transcriptions, additional noise may have been introduced into the system since pronunciations were not hand-designed. We see from the results in Table 3 that training on the automatically romanized transcriptions does not match the performance obtained by training on true romanizations, but it does decrease WER compared to only training on undiacritized script. The method could be improved by several additions which could not be implemented within the time frame of the workshop. Most importantly, acoustic information could be used when available. This could be achieved by listing all possible romanized forms as pronunciation variants for the script forms in the recognition lexicon, possibly weighted by prior probabilities computed by the methods used above, and letting the ASR training algorithm select the best option.

4. FACTORED LANGUAGE MODELING

Arabic has a rich morphology characterized by a high degree of affixation and interspersed vowel patterns and roots (sequences of three consonants) in word stems, as shown below:

fa + ta + D R u S + ina *(and you (f.) study)*

Here the stem consists of the root letters D-R-S with vowel *u* and

Word:	kan	mukalmAt	ziyAdaB
R:	NULL	klm	zUd
P:	NULL	mu-CACCAt	CiyACaB
M:	verb-past	noun f.pl.	noun f.sg.
S:	kAn	mukalmaB	ziyAdaB

Fig. 2. Decomposition of words into parallel morphological factors.

is surrounded by affixes *fa+*, *ta+* and *-ina*. As in other morphologically rich languages, the large number of possible word forms entails problems for robust language model estimation. It might therefore be preferable to use morpheme-like units instead of whole word forms as language modeling units. ASR systems based on morpheme sequences have been developed before (e.g. [4, 5]), where words were decomposed into linear sequences of morphemes which were then used as both acoustic as well as language modeling units. In most cases it was found that benefits gained from perplexity reduction were offset by the increased acoustic confusability due to the small size of morphemes. Moreover, standard trigram models over morphemes were found to be insufficient to model statistical relations across word boundaries. For these reasons we adopted a different approach. First, morphemes are used only in the language modeling component within an n-best list rescoring framework. N-best lists are generated in a first pass using trigram models over full word forms and are subsequently rescored using morpheme-based language models (LMs). Second, a given word w is not decomposed into a linear sequence of morphemes but into a bundle of n_w parallel components, such as stems (S), morph class (similar to affix specification) (M), patterns (P) and roots (R) (see Figure 2). If we assume that words can uniquely be defined in terms of, say, their roots, patterns, and morphological class, i.e. we have event equivalence such that $(W = w_i) \equiv (R = r_i, P = p_i, M = m_i)$. We can then define the trigram probability over words as follows

$$\begin{aligned} & P(w_i|w_{i-1}, w_{i-2}) \\ &= P(r_i, p_i, m_i|r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \\ &= P(r_i|p_i, m_i, r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \\ & \quad P(p_i|m_i, r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \\ & \quad P(m_i|r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \end{aligned} \quad (1)$$

Possible advantage of this *factored language model* (FLM)) are (a) more reliable estimation of the component probabilities, since more observations will be available for different combinations of morphemes as opposed to words, and (b) model simplification by eliminating superfluous conditioning variables in Equation 1, which can be done by a greedy search over possible combinations of conditioning variables. Moreover, this model can express dependencies across word boundaries and can integrate other (e.g. semantic) word features beyond morphological features. Since there is currently no generally available language modeling tool that directly supports the type of multi-stream input shown in Figure 2 we implemented a complete FLM module as an addition to the SRILM toolkit [6]. This module parses multi-stream input, correctly applies all major smoothing techniques both within and across streams, and also provides novel generalized backoff strategies (more details in [7]). Alternative, simpler ways of using morphological information are: (a) standard LMs for elements in individual streams, i.e. trigrams over roots or patterns only; (b) class-based models of the type

$$P(w_i|w_{i-1}, ..., w_{i-n+1}) \approx P(c_i|c_{i-1}, ..., c_{i-n+1})P(w_i|c_i) \quad (2)$$

where w is a word, n is the model order, and c denotes a class defined by a morphological component (e.g. stem); and (c) class-based models where not w_i but another morphological component (e.g. pattern, p_i) is predicted based on the current class c_i. We tested all four models (FLM and methods (a)-(c)) using the 4-way word decomposition shown in Figure 2, which was obtained using a combination of expert morphological knowledge and semi-automatic morphological analysis. First, words were decomposed into stems and morphological class based on information in the LDC ECA lexicon. Second, roots were extracted from stems using Darwish's Arabic morphological analyzer [8], and patterns were obtained by subtracting the root from the stem. Since this analyzer was developed for MSA, rather than ECA, the root and pattern output was errorful and less reliable than the hand-coded information from the LDC lexicon. The models were then used jointly with two word-based baseline LMs and acoustic scores in a log-linear combination scheme:

$$P(W|I) = \frac{1}{Z(I)} \prod_{i=1}^{k} S_k(W)^{\lambda_k} \quad (3)$$

where W is a word and I is a set of k knowledge sources whose scores $S_1(W), ..., S_k(W)$ are combined with weights $\lambda_1, ..., \lambda_k$. Z(I) is a normalization factor. The weights were optimized on the development set to minimize an objective function based on the smoothed word error count [9]. The two word-based models differed in the smoothing schemes used (Witten-Bell vs. Kneser-Ney) as well as their treatment of disfluencies and fillers, which were mapped to single classes in one model but were kept separate in the other. Table 4 shows the best results obtained by rescoring the n-best lists (n=100) of the romanized recognizer with the various LMs. In addition to using expert morphological classes for method (b), we also used automatically derived morphological components (based on work in [10]) and classes obtained by standard word clustering mechanisms (eg. as implemented in SRILM) for comparison. This did not lead to an improvement in performance (see Table 4) compared to using expert models, which may be due to the limited amount of training data we had available. Combination of all best models yielded a WER of 53.7%. An analysis of the trained weights showed that in addition to the word models, the stem and morph-class models consistently obtained the highest weights, which confirmed our assumption that these models contribute useful and complementary information.

Models	WER (%)
word1 (baseline)	55.1
word2	54.8
word1+word2	54.5
word1+word2+FLM	54.1
word2+M, (a)	54.1
word1+S+M, (b)	53.9
word1+word2+S, (c)	54.1
word1+word2+autom. morphology, (b)	54.3
word1+autom. clustering, (b)	54.2
combination of best models	**53.7**

Table 4. WERs on CH eval96 set obtained by combinations of different language models (S = stems, M = morph. class, P = patterns). See text for an explanation of methods (a-c).

5. USING OUT-OF-CORPUS TEXT DATA

In order to improve language modeling for CallHome we attempted to use a large corpus of MSA texts obtained from newspaper sources (such as An-Nahar and Al-Hayat) and TV broadcast transcriptions (Al-Jazeera). In contrast to the smaller CallHome model (trained on 150K words) the MSA model was trained on

over 300 million words. We tried various ways of integrating MSA data. First, we used a standard interpolation of both language models, with weights estimated on a held-out development set (0.03 for the MSA model vs. 0.97 for CH). Second, we combined this interpolation with a constrained backoff scheme which limited the amount by which the original n-gram probabilities in the CH model were allowed to change. Third, we tried a class-based model of the type shown in Equation 2 where classes were determined based on the MSA corpus (using the standard clustering procedure in the SRILM toolkit), and probabilities for class sequences and words given classes were estimated from the CH corpus. These methods reduced perplexity only insignificantly and did not lead to any improvements in word error rate. Finally, we used a more refined scheme of text selection, which was applied only to the TV transcriptions (9M words), which contained talk shows and interviews. Our goal was to select segments that were more conversational in nature. N-grams over part-of-speech (POS) sequences have previously been shown to be good indicators of conversational style [11]. For this reason we trained a statistical POS tagger on Treebank data for MSA (130K words) and the CH corpus. POS-based n-gram models were then estimated on each corpus separately. In order to select individual utterances we scored each utterance in the Al-Jazeera corpus with both n-grams and, similar to [11], computed the likelihood ratio as

$$Score(S_i) = log\frac{P(S_i|C)^{1/N_i}}{P(S_i|F)^{1/N_i}} \quad (4)$$

where S_i is the $i'th$ utterance in the corpus, N_i is the number of words in the utterance, C is the conversational language model (trained on CH), and F is the formal speech language model (trained on MSA data). We then ranked utterances by the resulting score, and selected the top k utterances such that the resulting number of words was roughly equal to that in the CH set. A new language model was trained on the resulting selection and was interpolated with the regular CH model. Although the selected set did contain utterances that were more informal in style (e.g. sports news as opposed to political news), it did not help in improving the word error rate on CH. Our final analysis showed that including all of the available language modeling text improved the trigram hit rate on CH only insignificantly; a plot of the likelihood ratio of the different data sets used for text selection (Figure 3) additionally shows that the statistics obtained from the TV transcriptions are very similar to that of the newspaper text. This indicates that MSA and ECA behave almost like two entirely different languages; standard model interpolation and data selection methods that are successful in other languages do not seem to help in this case.

6. CONCLUSIONS

The above results indicate useful future research directions for Arabic ASR. First,we showed that using phonetic information available in romanized as opposed to vowelless transcriptions significantly improves word error rate, and that it is possible to obtain improvements by using automatically romanized data. Automatic romanizations can be produced by bootstrapping a statistical predictor on a small amount of aligned script and romanized data. This has implications for future evaluation and data collection efforts: even though the evaluation standard may be a romanized representation of Arabic, data collection efforts should focus on collecting large amounts of undiacritized data and vowel-annotating only limited subsets. Second, we observed a small improvement by using morphologically-based language models. This trend needs to be verified on other Arabic ASR tasks and possibly more refined n-best lists or word lattices; however, an analysis of combination weights showed that models based on stems and morphological class information consistently received large weights, indicating that complementary information is inherent in those models. Finally, various methods of using MSA text data to improve the CallHome language model did not yield any improvement. Our analysis showed that MSA and ECA behave like two entirely different languages. Standard techniques for using out-of-corpus data that have proved successful in other languages fail in this case, indicating that significant WER improvements on the CH ECA tasks, and possibly future Arabic dialectal corpora, will be difficult to obtain without collecting appropriate amounts of language model data.

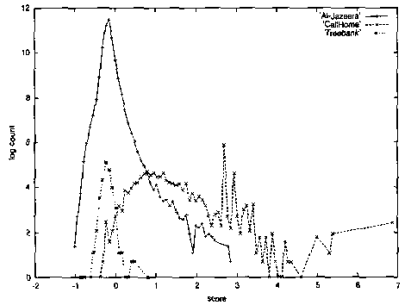

Fig. 3. Score (Eq. 4) distribution for CH (right-hand curve), Treebank (lower left-hand curve), and Al-Jazeera data (upper left-hand curve). The shapes of the Treebank and Al-Jazeera distributions are very similar and differ mainly in magnitude, whereas CH has a noticeably different distribution.

7. REFERENCES

[1] J. Billa et al., "Arabic speech and text in Tides OnTap," in *Proceedings of HLT*, 2002.

[2] J. Billa et al., "Audio indexing of broadcast news," in *Proceedings of ICASSP*, 2002.

[3] G. Zavagliakos et al., "The BNN Byblos 1997 large vocabulary conversational speech recognition system," in *Proceedings of ICASSP*, 1998.

[4] P. Geutner, "Using morphology towards better large vocabulary speech recognition systems," in *Proceedings of ICASSP*, 1995.

[5] P. Geutner K. Çarki and T. Schultz, "Turkish LVCSR: towards better speech recognition for agglutinative languages," in *Proceedings of ICASSP*, 2000.

[6] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings of ICSLP*, 2002.

[7] K. Kirchhoff et al., "Novel approaches to arabic speech recognition - final report from the JHU summer workshop 2002," Tech. Rep., John-Hopkins University, 2002.

[8] K. Darwish, "Building a shallow Arabic morphological analyser in one day," in *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, 2002.

[9] D. Vergyri, *Integration of Multiple Knowledge Sources in Speech Recognition Using Minimum Error Training*, Ph.D. thesis, Johns-Hopkins University, 2000.

[10] P. Schone, *Towards knowledge-free induction of machine-readable dictionaries*, Ph.D. thesis, University of Colorado at Boulder, 2001.

[11] R. Iyer, *Improving and predicting performance of statistical language models in sparse domains*, Ph.D. thesis, Boston University, 1998.