ARABIC BROADCAST NEWS TRANSCRIPTION USING A ONE MILLION WORD VOCALIZED VOCABULARY

Abdel. Messaoudi,[*†] *Jean-Luc Gauvain and Lori Lamel*
Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{abdel,gauvain,lamel}@limsi.fr

ABSTRACT

Recently it has been shown that modeling short vowels in Arabic can significantly improve performance even when producing a non-vocalized transcript. Since Arabic texts and audio transcripts are almost exclusively non-vocalized, the training methods have to overcome this missing data problem. For the acoustic models the procedure was bootstrapped with manually vocalized data and extended with semi-automatically vocalized data. In order to also capture the vowel information in the language model, a vocalized 4-gram language model trained on the audio transcripts was interpolated with the original 4-gram model trained on the (non-vocalized) written texts. Another challenge of the Arabic language is its large lexical variety. The out-of-vocabulary rate with a 65k word vocabulary is in the range of 4-8% (compared to under 1% for English). To address this problem a vocalized vocabulary containing over 1 million vocalized words, grouped into 200k word classes is used. This reduces the out-of-vocabulary rate to about 2%. The extended vocabulary and vocalized language model trained on the manually annotated data give a 1.2% absolute word error reduction on the DARPA RT04 development data. However, including the automatically vocalized transcripts in the language model reduces performance indicating that automatic vocalization needs to be improved.

1. INTRODUCTION

This paper describes some experiments aimed at addressing two challenges faced in transcribing broadcast news data in Modern Standard Arabic [9, 11]. These challenges are the explicit modeling of short vowels in the acoustic and language models, and dealing with the large lexical variety of Arabic. It has been recently shown that explicitly modeling short vowels improves recognition performance even when producing a non-vocalized transcript [1] over a grapheme-based approach [2] where only characters in the non-vocalized written form are modeled. We demonstrate that by building a very large vocalized vocabulary of more than 1.2 million words, and by using a language model including a vocalized component, the word error rate can be reduced significantly.

Arabic is a strongly consonantal language with nominally only three vowels, each of which has a long and short form. The vowels and gemination marks are generally not indicated in written texts. It is a highly inflected language, with many different word forms for a given root, produced by appending articles ("the, and, to, from, with, ...") to the word beginning and possessives ("ours, theirs, ...") on the word end. There are typically several possible (generally semantically linked) vocalizations for a given written word, which are spoken. The word-final vowel varies as a function of the word context, and this final vowel or vowel-'n' sequence is often not pronounced.

Thus one of the challenges of explicitly modeling vowels in Arabic is to obtain vocalized resources, or to develop efficient ways to use non-vocalized data [12]. It is often necessary to understand the text in order to know how to vowelize and pronounce it correctly. In [9] we investigated using the Buckwalter Arabic Morphological Analyzer [1] to propose possible multiple vocalized word forms, and then used a speech recognizer to automatically select the most appropriate one during acoustic model training. In this paper we extend this approach, previously used for acoustic modeling, to language modeling.

2. PRONUNCIATION LEXICON

The pronunciation dictionary was developed in several steps. First, all distinct words in the transcripts of the audio training data are included in the pronunciation dictionary [11]. As described in [9], for the portion of the data transcribed with vowels, all vocalized forms of a given non-vocalized script are associated with that entry. Each entry can be thought of as a word class, containing all observed vocalized forms of the word. Given the limited amount of vocalized training data, there are many possible vocalized word forms that are never observed. The Buckwalter Arabic Morphological Analyzer (version 1) was used to complete the lexicon, proposing forms that did not occur in the training data. The Buckwalter analyzer was also used to propose vocalized forms for the part of the audio training data

[*†] Visiting scientist from the Vecsys Company.

[1] T. Buckwalter, http://www.qamus.org/morphology.htm

subject	/u/ (damma)
direct object	/a/ (fatha)
indirect object	/i/ (kasra)
indefinite	/n/ (tanwin)

Table 1: Rules to add vowels to Arabic words.

for which only non-vocalized transcripts were available. Vocalized forms for proper names and technical terms (about 1500) were derived manually for the 65k dictionary.

Since the morphological analyzer (v1.0) does not produce all possible forms for the final vowel, a set of rules were developed to automatically generate these alternate forms. The final vowel varies according to the grammatical role of the word as summarized in Table 1. If the word begins with the article "al", the permissible final vowels are /i,a,u/. If not, the word could be indefinite, and the tanwin forms (doubling of the vowel) are proposed. Other rules limit the possible vowels if certain prefixes ("bi", "li", "biAal", ...) or word ending (Alif, Alif maksoura or ta marbouta) are present. Other rules depend on the gender and number of the word, and the presence of a suffix indicating a possessive pronoun. A set of rules apply to the hamza. While the stable hamza is always pronounced, the unstable hamza is often omitted except at the beginning of a sentence or after a pause. Therefore, for the unstable hamza two vocalized forms are generated. Another rule treats the marking of gemination for solar consonants following the article "al". In total there are about 30 rules used to generate alternate vocalized word forms. The number of rules will likely be reduced with the new release of the Buckwalter analyzer (version 2).

Letter to sound conversion in Modern Standard Arabic is quite straightforward when starting from vocalized texts. A grapheme-to-phoneme conversion tool was developed using a set of 37 phonemes and three non-linguistic units (silence/noise, hesitation, breath). The phonemes include the 28 Arabic consonants (including the emphatic consonants and the hamza), 3 foreign consonants (/p,v,g/), and 6 vowels (short and long /i/, /a/, /u/). Alternate pronunciation variants are included to allow for differences in how MSA is spoken in Egypt (the /J/ is pronounced /g/) in contrast to other regions.

The 65k word vocabulary has 440k different vocalized forms with 530k phone transcriptions. The 200k word vocabulary has 1.2 M distinct vocalized forms and 1.4 M pronunciations.

3. LANGUAGE MODELS

A 65k and 200k non-vocalized word language models have been built by interpolating 11 backoff 4-gram models [9] trained on the 4 sources of the LDC Arabic Gigaword corpus [8] (390M words), on 6 sources collected from the Internet (204M words), and on the manual transcriptions of the acoustic data used to trained the acoustic models (1.1M words). The texts were preprocessed to remove undesirable material, transliterated using a slightly extended version of Buckwalter algorithm,[2] and normalized in order to better approximate a spoken form [4, 9]. After processing there were a total of 600 million words, of which 2.2 M are distinct. The out-of-vocabulary (OOV) rates for the language models are 4.4% and 2.0% respectively. The language model interpolation weights were tuned to minimize the perplexity on a set of development shows from November 2003 shared by BBN.

A 1.2 million word vocalized word language model has then been built by interpolating the non-vocalized LM and a vocalized LM trained on the vocalized part of the manual transcriptions. This data is comprised of a total of 580k words with 85k different vocalized forms and 50k distinct non-vocalized forms. As described above the vocalized vocabulary has been obtained by semi-automatically generating all possible vocalized forms for the 200k non-vocalized word vocabulary. The vocalized n-gram probabilities $P(v_i|v_{i-1},...)$ are estimated in the following way (v_i and w_i are respectively the vocalized and non-vocalized forms of ith word):

$$\begin{aligned} P(v_i|v_{i-1},...) &= \alpha P_a(v_i|v_{i-1},...) + \\ &\quad (1-\alpha)P_v(v_i|w_i)P_t(w_i|w_{i-1},...) \end{aligned}$$

where P_a is the vocalized LM trained only on the vocalized part of the acoustic data, P_v is trained on all the acoustic data after Viterbi alignment, and P_t is the standard non vocalized LM trained on all of the data described above. Adding the automatically vowelized transcripts to the data used to estimate the vocalized LM not improve performace. Independent of whether a vocalized or non vocalized language model is used, the decoder outputs a non vocalized transcription. When using the vocalized LM, the posterior probabilities of the vocalized forms corresponding to the same non-vocalized word are summed to compute the word posterior probabilities. This is the same as what is done for consensus decoding with alternate pronunciations.

4. RECOGNITION SYSTEM OVERVIEW

The LIMSI broadcast news transcription system has two main components, an audio partitioner and a word recognizer. Data partitioning is based on an audio stream mixture model [3, 4], and serves to divide the continuous stream of acoustic data into homogeneous segments, associating cluster, gender and labels with each non-overlapping segment. For each speech segment, the word recognizer determines the sequence of words in the segment, associating start and end times and an optional confidence measure with

[2] http://www.qamus.org/transliteration.htm

each word. The recognizer makes use of continuous density HMMs for acoustic modeling and n-gram statistics for language modeling. Each context-dependent phone model is a tied-state left-to-right CD-HMM with Gaussian mixture observation densities where the tied states are obtained by means of a decision tree.

Word recognition is performed in three passes, where each decoding pass generates a word lattice which is expanded with a 4-gram LM. Then the posterior probabilities of the lattice edges are estimated using the forward-backward algorithm and the 4-gram lattice is converted to a confusion network with posterior probabilities by iteratively merging lattice vertices and splitting lattice edges until a linear graph is obtained. This last step gives comparable results to the edge clustering algorithm proposed in [10]. The words with the highest posterior in each confusion set are hypothesized.

The first decoding pass generates initial hypotheses which are then used for cluster-based acoustic model adaptation. This is done via one pass (less than 1xRT) cross-word trigram decoding with gender-specific sets of position-dependent triphones (5700 tied states) and a trigram language model (38M trigrams and 15M bigrams). Band-limited acoustic models are used for the telephone speech segments. The trigram lattices are rescored with a 4-gram language models. These hypothesis are used to carry out unsupervised acoustic model adaptation is performed for each segment cluster using the MLLR technique [7] with only one regression class. The lattice is generated for each segment using a bigram LM and position-dependent triphones with 11500 tied states (32 Gaussians per state). The word graph generated in the second decoding pass is rescored after carrying out unsupervised MLLR acoustic model adaptation using two regression classes.

The acoustic training data is comprised of about 150 hours of radio and television broadcast news data from a variety of sources. About half of the data were manually transcribed with short vowels and other diacritic marks [11], so as to enable accurate modeling of the short vowels. Vocalized transcripts were not available for the remaining audio data, from the TDT4 and FBIS corpora [9]. The time-aligned segmented transcripts, shared with us by BBN, were derived from the associated closed-captions and commercial transcripts. Training on these data was semi-supervised, allowing the recognizer to choose the preferred form from the vocalized pronunciations associated with the non-vocalized written in the lexicon.

The acoustic models are context-dependent, 3-state left-to-right hidden Markov models with Gaussian mixture. Two sets of gender-dependent, position-dependent triphones are estimated using MAP adaptation of SI seed models for wide-band and telephone band speech [5]. The triphone-based context-dependent phone models are word-independent but word position-dependent. The first decoding pass uses a small set of acoustic models with about 5700 contexts and tied states. A larger set of acoustic models, used in the second and third passes, cover about 15800 phone contexts represented with a total of 11500 states, and 32 Gaussians per state. State-tying is carried out via divisive decision tree clustering, constructing one tree for each state position of each phone so as to maximize the likelihood of the training data using single Gaussian state models, penalized by the number of tied-states [4]. A set of 152 questions concern the phone position, the distinctive features (and identities) of the phone and the neighboring phones.

The training data were also used to build Gaussian mixture models with 2048 components which are used for acoustic model adaptation in the first decoding pass.

Language model	*WER*	*OOV*
65k LM	16.0%	4.4%
200k LM	15.2%	2.0%
1.2M vocalized LM	14.8%	2.0%

Table 2: Word error rates on the NIST RT04 development data with the 65k and 200k non-vocalized word LM and with the 1.2M vocalized LM.

5. EXPERIMENTAL RESULTS

Recognition experiments were carried out using the development data for the NIST RT-04 evaluation (www.nist.gov/speech/tests/rt). The development data consist of 3 shows broadcasts recorded at the end of November 2003 from Al-Jazeera and Dubai TV, for which reference transcriptions were shared by BBN. (Since normalization of the reference transcriptions can have a large influence on the error rate, mapping rules were exchanged with BBN.)

The non-vocalized word error rate of with the 65k vocabulary is 16.0%, with an OOV rate of 4.4%. Extending the vocabulary to 200k non-vocalized entries, cuts the OOV rate in half, and results in a 5% relative reduction in the word error rate to 15.2%. Interpolating the vocalized LM component with the non-vocalized 4-gram LM further reduces the word error rate to 14.8%. This is a large gain given the limited quantity of vocalized LM training data.

Since including the audio training with implicitly vocalized transcripts was effective for acoustic modeling, we therefore decided to try to use the approximately 520k words of automatically vocalized transcripts to increase the amount of vocalized LM training data. The decoder is used to produce the vocalized orthographic form associated with each word hypothesis (instead of the non-vocalized word class). However, the vocalized LM trained on the combined manually and automatically vocalized data performed less well than the one trained on only the manually annotated data. This led us to conclude that there are too many errors on the

Substitution	*%of occurrences*
'alyawm → lyawm	55.6%
min → man	7.7%
'allah → llah	81.8%
'arra'Is → rra'Is	71.9%
'anna → 'ana	14.3%
'assAbiq → ssAbiq	86.0%
fI → fiyya	2.5%
mac → maca	55.4%
'addawliyya → ddawliyya	94.3%
l'amIrikiyya → l'amirIkiyya	52.5%

Table 3: Most frequent alignment substitution errors.

vowels, and prompted us to look into the most frequent error types.

In order to study the capability of the system to add vowels to a non-vocalized reference transcript, a set of vocalized test data from eight audio sources was used [11]. The audio data were aligned with the non-vocalized reference via the recognition lexicon, which included all permissible vocalizations for each entry. When vowel errors are counted, the word error rate is approximately doubled, however most (3/4) of the errors are on the final vowel. In order to avoid normalization problems, the vowel error rate was computed at the phone level. We estimate the vowel error to be about 30% for the final vowel, and about 10% in other word positions.

About 25% of the errors can be attributed to the unstable hamza. The 10 most frequent substitution errors are shown in Table 3 along with the percentage of occurrences of the reference word that has this substitution. Six of the most frequent errors have the elision of the word initial hamza, of which 5 also elide the short vowel /a/. The last entry is a confusion between long and short /i/ in the word "America".

6. CONCLUSIONS

This paper has reported on our recent development work on transcribing Modern Standard Arabic broadcast news data and on explicitly modeling short vowels in the acoustic and language models, even though these are removed prior to scoring. In order to be able to make use of non-vocalized audio and textual resources, the recognition lexicon entries are word-classes which regroup all derived vocalized forms along with the associated phonetic forms. To address the large lexical variety of Arabic and reduce the out-of-vocabulary rate, the recognition vocabulary has been extended from 65k to 200k word-classes (with over 1 million vocalized words). The 65k vocabulary contains 529k phone transcriptions, and the 200k has over 1.4M pronunciations. The explicit internal representation of vocalized word forms in the lexicon may be useful to provide an automatic (or semi-automatic) method to vocalize transcripts. The audio data without explicit vowels has been successfully used for acoustic modeling which can reduce the cost and ease of data transcription. However our attempts to also use the automatically derived vocalized forms for language modeling was unsuccessful, indicating that the vowel error rate (in particular for the final vowel) is too high. Building a very large vocalized vocabulary of more than 1.4 million words, and by using a language model including a vocalized component, significantly reduced the word error rate.

REFERENCES

[1] M. Afify, L. Nguyen, B. Xiang, S. Abdou and J. Makhoul, "Recent Progress in Arabic Broadcast News Transcription at BBN" *Eurospeech'05*, 1637-1640, Lisbon, Sep. 2005.

[2] J. Billa, N. Noamany, A. Srivastava, D. Liu, R. Stone, J. Xu, J. Makhoul, F. Kubala, "Audio Indexing of Arabic Broadcast News," *ICASSP'02*, **1**:5-8, Apr 2002.

[3] J.L. Gauvain, L. Lamel, G. Adda, "Partitioning and Transcription of Broadcast News Data," *ICSLP'98*, **5**:1335-1338, Dec 1998.

[4] J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, **37**(1-2):89-108, May 2002.

[5] J.L. Gauvain, C.H. Lee, "Maximum A Posteriori for Multivariate Gaussian Mixture Observation of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, **2**(2):291-298, Apr 1994.

[6] L. Lamel, J.L. Gauvain, "Automatic Processing of Broadcast Audio in Multiple Languages," *Eusipco'02*, Sep 2002.

[7] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, **9**(2):171-185, 1995.

[8] Linguistic Data Consortium. The Arabic Gigaword corpus (LDC2003T12), 2003.

[9] A. Messaoudi, L. Lamel and J.L. Gauvain, "Modeling Vowels for Arabic BN Transcription," *Eurospeech'05*, 1633-1636, Lisbon, Sep. 2005.

[10] L. Mangu, E. Brill, A. Stolke, "Finding Consensus Among Words: Lattice-Based Word Error Minimization," *Eurospeech'99*, 495-498, Sep 1999.

[11] A. Messaoudi, L. Lamel, J.L. Gauvain, "Transcription of Arabic Broadcast News," *ICSLP'04*, Oct 2004.

[12] D. Vergyri and K. Kirchhoff, "Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition", COLING Workshop on Arabic-script Based Languages, Geneva, Switzerland, 2004