INTERSPEECH 2009 BRIGHTON

Lexical and Phonetic Modeling for Arabic Automatic Speech Recognition

Long Nguyen, Tim Ng, Kham Nguyen[1], Rabih Zbib[2], and John Makhoul

BBN Technologies, 10 Moulton Street, Cambridge, MA 02138, USA
[1]Northeastern University, Boston, Massachusetts, USA
[2]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
{ln,tng,knguyen,rzbib,makhoul}@bbn.com

Abstract

In this paper, we describe the use of either words or morphemes as lexical modeling units and the use of either graphemes or phonemes as *phonetic* modeling units for Arabic automatic speech recognition (ASR). We designed four Arabic ASR systems: two word-based systems and two morpheme-based systems. Experimental results using these four systems show that they have comparable state-of-the-art performance individually, but the more sophisticated morpheme-based system tends to be the best. However, they seem to complement each other quite well within the ROVER system combination framework to produce substantially-improved combined results.

Index Terms: Arabic speech recognition, morphological decomposition, lexical unit, phonetic unit, system combination

10.21437/Interspeech.2009-244

1. Introduction

Arabic presents two primary difficulties to automatic speech recognition (ASR) systems, both due to the Arabic writing system. The first difficulty is the morphological complexity that arises from the attachment of affixes, which renders recognition vocabularies very large if one wants to minimize out-of-vocabulary (OOV) words. For example, a 65K Arabic vocabulary typically has an OOV rate of 5%, compared to 0.5% for the same size English vocabulary. The second difficulty arises due to the fact that the short vowels are typically not written. Both difficulties are partially overcome by the use of automatic morphological analyzers, which also produce vocalizations, but the resulting analysis and vocalization can be errorful at times.

In this paper, we describe the use of either words or morphemes as lexical units and the use of either graphemes or phonemes as *phonetic* units in Arabic ASR to handle the two key problems mentioned above. Specifically, we designed four Arabic ASR systems: two word-based systems and two morpheme-based systems. In the first word-based system, each word is modeled by one or more sequences of phonemes of its phonetic pronunciations. In the second word-based system, each word is modeled by exactly one sequence of letters of its spelling. The third and fourth systems use morphemes as recognition units. Morphemes are determined by either a simple morphological decomposition using a small set of affixes and a few rules or by an elaborate linguistically-driven morphological analyzer. Each morpheme is modeled by sequences of phonemes of its pronunciations derived from the corresponding word's pronunciations during the decomposition process.

Experimental results using these four systems (developed for the GALE Phase 3 evaluation) show that they have comparable state-of-the-art performance individually. However, they all seem to complement each other quite well such that the combination of the four systems using ROVER provides substantial improvement in performance when compared to each single system.

2. System Description

At the core of each ASR system is the Byblos multi-pass recognizer. Various acoustic and language models at different levels of sophistication are deployed at different passes and/or stages.

2.1. Recognizer

The Byblos multi-pass recognizer [1] first performs a fast match of the data to produce scores for numerous word endings using a coarse state-tied-mixture (STM) acoustic model (AM) and a bigram language model (LM). Next, a state-clustered tied-mixture (SCTM) AM and an approximate trigram LM are used to generate lattices. Lattices are then re-scored using a cross-word SCTM AM and a 4-gram LM. The best path of the re-scored lattice is the recognition result. In other words, the decoding process is a three-step sequence (fast-match, lattice generation, and lattice rescoring) with finer-detailed models being used on narrower search space at later steps [2].

The decoding process is repeated three times. First, speaker-independent (SI) acoustic models are used in the decoding to generate hypotheses for unsupervised adaptation. Then, the decoding is repeated but with speaker-adaptively-trained (SAT) acoustic models that have been adapted to the hypotheses generated in the first stage. The last decoding is similar to the second but acoustic models are adapted to the second stage's hypotheses using a larger number of regression classes.

2.2. Acoustic Model Training

The typical procedure for training acoustic models at BBN can be logically grouped into these four sequential stages.

Front-end Processing: 14-dimensional Perceptual Linear Predictive [3] cepstral coefficients are extracted from the overlapping frames of audio data with a frame rate of 10ms. Cepstral mean subtraction is applied for normalization. The normalized energy is used as the 15^{th} component. Nine successive 15-component frames centered at the current frame are concatenated and then reduced to a 60-dimensional feature vector using Linear Discriminant Analysis (LDA) and decorrelated using Maximum Likelihood Linear Transformation (MLLT) [4]. The dimension reduction is applied differently for SI and SAT models.

ML-SI Training: The SI AMs are trained using the Maximum-Likelihood (ML) criterion. Feature dimension reduction is done via global LDA and MLLT transformations estimated from and applied to all training data. These models are to be used in the SI decoding stage.

ML-SAT Training: The SAT AMs are then trained using also the ML criterion. Each 15-dimensional feature vector is transformed using a speaker-cluster-dependent Constrained Maximum Likelihood Linear Regression (CMLLR) transform [5] before the concatenation of nine successive frames for dimension reduction. Both LDA and MLLT transformations are speaker-cluster-specific at this stage. The resulting 60-dimensional feature vectors are further transformed using new CMLLR transformations. These models are used only in the adapted decoding stages.

MPFE Training: In the last stage of acoustic model training, all training data is decoded using the ML models to generate lattices. Then a new set of AMs, both SI and SAT, are estimated using these lattices under the Minimum Phoneme-Frame Error criterion [6].

The total amount of acoustic training data used in this effort is about 1400 hours selected through light supervision [7] of all data available to the GALE community. Specifically, the data consist of the following corpora: BBN-FBIS (43hr), TDT4 (67hr), Iraqi Dialect (50hr), GALE Year1 (144hr), Phase2 (603hr), and Phase3 (526hr).

2.3. Language Model Training

The language models were estimated through interpolation of several sub N-Gram models trained on disjoint subsets of the 1.7-billion words of Arabic text data available to the GALE community as of May 2008. Each individual sub N-Gram model was trained using the modified Kneser-Ney smoothing technique.

3. Implementation

Arabic Text Normalization: Due to writing conventions in Arabic, sometimes the same word has different written forms. This is especially true for words that start with the letter "hamza" (corresponding to the glotal stop). So, for consistency in lexical representation for ASR, we map all different forms of "hamza" ("<", ">", and "|" using Buckwalter transliteration scheme) at the beginning of the word, or after the common Arabic prefixes "Al" and "w", to "alif" ("A"). Also, for certain frequent words, we map the "alif maksura" ("Y") at the end of the word to "yeh" ("y") or vice versa. However, for scoring of ASR output, all forms of "hamza" are equated to "alif" without the restriction above.

Master Phonetic Dictionary: As described in [8], [9], and [10], we constructed phonetic dictionaries for Arabic by using the Buckwalter morphological analyzer [11] and manually-vocalized corpora. Minor changes to this procedure include the addition of extra affixes used in major Arabic dialects to the list of Modern Standard Arabic (MSA) affixes used in the Buckwalter analyzer. We also added all vocalizations found in the manually-vocalized corpora (LDC's Arabic TreeBanks, LDC's Iraqi dialect lexicon, and the 45-hour BBN-FBIS corpus). As a result, our Arabic master phonetic dictionary, as of now, consists of about 1.2 million words, each with about 3.76 pronunciations on average. This is the master dictionary from which we derive phonetic pronunciations for all phonetic word-based and morpheme-based Arabic ASR systems.

3.1. Phonetic System (P)

The design of our word-based phonetic system (P) is a straightforward implementation of a typical ASR system. The recognition units are Arabic words and each word is modeled by one or more sequences of phonemes of its pronunciation(s). The largest AM (i.e. the cross-word SCTM model) has about 220K cross-word quinphone *states* sharing 7K sets of Gaussians (or codebooks). Note that we use two-level Gaussian mixture models with several different states having separate sets of mixture weights but sharing the same set of Gaussians (or codebooks, as described in details in [2]). The total number of Gaussians in this model is about 900K. The exact values for all four systems can be found in Table 1.

Sys.	#states	#codebooks	#Gaussians
P	220694	6882	889146
G	212066	6580	867707
M1	222506	6829	883117
M2	220561	6854	885518

Table 1: AM comparison: all four systems use about the same number of model parameters.

The recognition vocabulary used in this phonetic system consists of 390K words. Each word has 4.03 pronunciations on average. This is a subset of the 490K words that occur at least 30 times in the language model training data or at least 3 times in the transcripts of the acoustic training data. That means 100K words in the list of the 490K most frequent words do not exist in our 1.2M-word master dictionary because they could not be analyzed by the Buckwalter analyzer and they do not occur in the manually-vocalized corpora. This system's language models consists of 137M trigrams (pruned based on entropy) used for decoding and 585M (unpruned) four-grams used for lattice rescoring. The OOV rate of this vocabulary and the size of the LMs of this system are compared against those of other systems in Tables 2 and 3.

Sys.	vocab.	pron.	#3-grams	#4-grams
P	390K	4.03	136,767,904	585,331,357
G	490K	1.00	140,915,143	617,448,176
M1	289K	4.43	137,586,807	565,474,893
M2	284K	3.69	142,068,884	563,803,750

Table 2: LM comparison: all four systems have about the same number of n-grams even though the size of the morpheme vocabularies is substantially smaller than that of the word vocabularies.

3.2. Graphemic System (G)

The word-based graphemic system (G) is similar to the phonetic system except that it doesn't use phonetic pronunciations and it has a much larger recognition vocabulary. The recognition units are also Arabic words but each word is modeled by exactly one sequence of letters of its spelling. The largest AM of this system has about 210K states sharing 7K sets of Gaussians for a total of about 900K Gaussians (as shown in row G of Table 1). Since this system doesn't depend on the real phonetic dictionary, its recognition vocabulary uses all 490K most frequent words. Its LM consists of 140M (pruned) trigrams and 617M (unpruned) four-grams.

3.3. Morphemic System 1 (M1)

The morphemic system 1 (M1) is a morpheme-based phonetic system. Morphemes are determined by a simple morphological decomposition of the words without their context using a small set

of affixes and a few rules. More details on various morphological decomposition strategies and their effect to ASR performance for morphemic systems developed in the past can be found in [9]. In the current M1 system, the decomposition process uses 12 prefixes and 34 suffixes. The prefixes are: *Al, bAl, fAl, kAl, ll, wAl, b, f, k, l, s, w*, and the suffixes are: *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*. The process also utilizes a list of 128K most frequent decomposable words that should not be decomposed (hereafter referred to as a *blacklist*) since it has been shown that this really improved recognition performance. For a candidate word to be decomposed, it has to satisfy the following five conditions: (1) it doesn't belong to the blacklist, (2) it consists of at least one of the pre-determined affixes, (3) its decomposed affixes' pronunciations match the pre-determined pronunciations, (4) its stem must exist in the master dictionary, and (5) its stem must be at least two letters long.

As shown in Table 1, the acoustic models of this system were designed to have about the same number of system parameters as all other systems. However, the recognition units, i.e. the *morphemes*, in this system consist of 289K morphemes obtained through the decomposition of all 490K most frequent words, including its blacklist, and all words occurring in the transcripts of the acoustic training data. On average, each morpheme has about 4.43 pronunciations. The LM also has about the same number of n-grams used in the other systems, as shown in Table 2.

3.4. Morphemic System 2 (M2)

The morphemic system 2 (M2) is another morpheme-based phonetic system, but the morphemes in this case were determined by a more sophisticated decomposition process. Briefly, the process consists of two steps. First, we use Sakhr's Arabic morphological analyzer[1] to decompose all AM and LM training data (of about two billion words). Each word, if occurring more than once, can be decomposed into different sequences of morphemes depending on its different contexts. Note that each word instance is decomposed into exactly one sequence of morphemes of the form *[prefix] + stem + [suffix],* where either the prefix or the suffix or both can be missing. Then, we collect all the decomposed forms of the 490K most frequent words and all words occurring in the acoustic training data and pass them through the final decomposition process. Similar to the construction of the recognition units in system M1, we also utilize a blacklist of the 128K most frequent decomposable words. The decomposable words are finally decomposed as follows:

- if the word is in the blacklist, keep unchanged;
- else if no *prefix*, decompose into *stem* and *suffix*;
- else if no *suffix*, decompose into *prefix* and *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*.

[1]The authors would like to thank Sakhr for the use of their proprietary Arabic morphological analysis software for this study. Sakhr is a member of the BBN-led GALE AGILE team.

This decomposition process produced about 284K morphemes to be used in the recognition vocabulary in this system. On average, each morpheme has about 3.69 pronunciations. Both the AMs and LMs of this system have about the same sizes compared to those of other systems, as shown in Tables 1 and 2. A full description of this system and a comparison to the M1 morphemic system can be found in [10].

4. Experimental Results

Development Test Sets: To support the research and development of these four systems, we used five Arabic test sets. The first two test sets, **at6** and **ad6**, were constructed at BBN at the start of the GALE program. Each is about six hours long, and uses Arabic broadcast programs aired in November 2005 and January 2006. These two test sets contain long segments of broadcast *stretches* of about ten to twenty minutes long. The remaining three test sets are development or evaluation test sets designed by LDC/NIST to be used by all GALE participants. Each of these three test sets is about three hours long. They consist of Arabic broadcast programs aired in November 2006 (**dev07**), December 2006 (**eval07**), and May 2007 (**dev08**). These three test sets contain only short *snippets* ranging from two to four minutes long (to facilitate the research and evaluation of the down-stream machine translation task of the GALE program). Each test set includes both *broadcast news* and *broadcast conversations* with almost equal amounts.

4.1. OOV Comparison

As shown in Table 3, both word-based systems (P and G) have pretty high OOV rate for all the test sets even though they use rather large vocabulary (400-500K words). Except for **dev08**, the OOV rates are 3% or 4% for all other test sets. The graphemic system (G) has a bit lower OOV rate since it uses a larger vocabulary. In contrast to using words as recognition units, both morphemic systems (M1 and M2) have much lower OOV rate even though their vocabulary sizes are substantially smaller. It is interesting to point out that the vocabulary of the M2 system has similar OOV rates across all test sets. Note that we scaled the morpheme OOV rate of the morpheme vocabulary by the decomposition rate (number of morphemes after decomposition divided by number of words before decomposition) to make it comparable to the standard word OOV rate.

Sys.	#units	at6	ad6	dev07	eval07	dev08
P	390K	4.02	3.39	4.36	2.88	1.44
G	490K	3.43	2.86	3.78	2.07	0.84
M1	289K	1.31	1.20	2.82	1.89	0.94
M2	284K	0.66	0.68	0.81	0.66	0.56

Table 3: OOV rate comparison: morpheme-based vocabularies have better coverage than word-based vocabularies for Arabic.

4.2. Individual Results

Table 4 shows the WER of the four systems on the five test sets. Overall, systems using morphemes as recognition units perform better than systems using words as recognition units. Using morphemes derived from the linguistically-driven approach in M2 produces the lowest WER among the four systems. Between the two morphemic systems, M2 tends to be more robust in terms of both

OOV rate and WER. M2 is clearly the best system on the two BBN-constructed test sets (at6 and ad6). Between the two word-based systems, the phonetic system is better than the graphemic system for the three GALE test sets (dev07, eval07, and dev08). However, the graphemic system is better than the phonetic system on the two BBN test sets (at6, and ad6). It's important to point out that the two BBN test sets are much more difficult than the three LDC/NIST test sets based on WER.

These four sets of results seem to correlate well with the complexity of the design of an Arabic ASR system. The simplest system has the highest WER (G) and the most sophisticated system (M2) has the lowest WER among the four systems. This finding would probably provide sufficient information for a designer of an Arabic ASR system to decide the type of recognition units to use to achieve the best trade-off required by his/her application. For example, even though it has the worst performance among the four systems, the word-based graphemic approach is clearly the simplest and easiest to implement, based on the fact that it doesn't require a phonetic dictionary. Furthermore, each word would have exactly only one *pronunciation* and, consequently, the recognizer would run faster in comparison to phonetic systems that have four pronunciations per word on average. However, if accuracy is the top requirement, the designer must be prepared to construct a phonetic dictionary with good quality and coverage. Also, a decent morphological analyzer that can decompose words within context into morphemes is required.

Sys.	at6	ad6	dev07	eval07	dev08
P	18.8	16.9	10.6	11.6	12.1
G	18.5	16.7	11.6	12.2	12.5
M1	18.1	17.1	10.3	11.1	11.6
M2	17.6	16.3	10.2	10.8	11.8

Table 4: Individual system results: the simplest system (G) has the overall highest WER and the most sophisticated system (M2) has the lowest WER.

4.3. System Combination Results

In addition to the scientific curiosity to search for the optimal recognition unit for Arabic ASR, another goal of this work is to have diverse but complementary systems to be used within the ROVER system combination framework. As shown in Table 5, the combination of all four systems (P+G+M1+M2) provides the lowest WER for all test sets. The relative reduction in WER for all five test sets in comparison the the best system (M2) is from five to ten percent. Various combinations of two or three systems, starting with the phonetic system (P), are shown in the top two blocks of Table 5. Generally, combining two systems produces about zero to three percent relative reduction in WER. Combining three systems increases the relative WER reduction further to about three to six percent.

ROVER	at6	ad6	dev07	eval07	dev08
P+G	17.4	15.8	10.5	10.9	11.6
P+M1	17.7	16.4	10.1	10.9	11.4
P+M2	17.3	15.9	10.2	10.7	11.5
P+G+M1	16.9	15.6	9.9	10.6	11.0
P+G+M2	16.7	15.4	9.8	10.4	11.0
P+M1+M2	17.2	15.8	9.8	10.5	11.1
P+G+M1+M2	16.6	15.3	9.7	10.3	10.8

Table 5: Various system combination results: combining two systems (P+*) produces 0-3% in WER reduction relative the best single system (M2); combining three systems (P+*+*), 3-6%; and the lowest WER obtained by combing all four systems.

5. Conclusion

We have described two possible types of lexical recognition units, either words or morphemes, and the use of either graphemes or phonemes as *phonetic* modeling units for Arabic automatic speech recognition. The study involves implementation of four different Arabic ASR systems. The results of these four systems show that they all have comparable state-of-the-art performance. They also show a good correlation to the complexity of the design of an Arabic ASR system: the simplest system, the word-based graphemic system, has the highest WER and the most sophisticated system, the linguistically-driven morphemic system, has the lowest WER. In addition, even though all four systems use the same underlying ASR technology, the use of different recognition units seem to be quite complementary to each other such that the combination of the four systems produces substantially-improved combined results.

6. References

1. L. Nguyen and R. Schwartz, "Efficient 2-pass N-Best decoder," *Proc. EuroSpeech*, Rhodes, Greece, Sep. 1997, pp. 167-170.
2. L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, R. Schwartz and J. Makhoul, "Progress in transcription of broadcast news using Byblos," *Speech Communication*, 38, pp. 213-230, 2002.
3. H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, **87**(4):1738-1752, April 1990.
4. B. Zhang, S. Matsoukas, and R. Schwartz, "Long span features and minimum phoneme error heteroscedastic linear discriminant analysis," *Proc. EARS RT-04 Workshop*, New York, Sep. 2004.
5. M. J. F. Gales, "Maximum Likelihood Linear Transformation for HMM-based Speech Recognition," *Tech. Report CUED/F-INFENG/TR291*, Cambridge University Engineering Dept., 1997.
6. J. Zheng and A. Stolke, "Improved discriminative training using phone lattices," *Proc. InterSpeech*, Lisbon, Portugal, Sep. 2005.
7. L. Nguyen and B. Xiang, "Light supervision in acoustic model training," *ICASSP'04*, Montreal, May 2004.
8. M. Afify, L. Nguyen, B. Xiang, S. Abdou, and J. Makhoul, "Recent progress in Arabic Broadcast News Transcription at BBN," *Proc. InterSpeech*, Lisbon, Portugal, Sep. 2005.
9. B. Xiang, K. Nguyen, L. Nguyen, R. Schwartz, and J. Makhoul, "Morphological Decomposition for Arabic Broadcast News Transcription," *ICASSP'06*, Toulouse, France, May 2006.
10. T. Ng, K. Nguyen, R. Zbib, and L. Nguyen, "Improved Morphological Decomposition for Arabic Broadcast News Transcription," *ICASSP'09*, Taipei, Taiwan, Apr 2009.
11. T. Buckwalter, *http://www.qamus.org/morphology.htm*