interspeech 2008 incorporating SST 08

Investigating Morphological Decomposition for Transcription of Arabic Broadcast News and Broadcast Conversation Data *

Lori Lamel, Abdel. Messaoudi, Jean-Luc Gauvain

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, France

1. ABSTRACT

One of the challenges of Arabic speech recognition is to deal with the huge lexical variety. Morphological decomposition has been proposed to address this problem by increasing lexical coverage, thereby reducing errors that are due to words that are unknown to the system. In our previous attempts to develop an Arabic speech-to-text (STT) transcription system with morphological decomposition, an increase in word error rate of about 2% absolute was observed relative to a comparable word based system. Based on an error analysis and a comparison of our approach with that of other sites, two modifications were made. The first modification was to not decompose the most frequent words; and the second to not decompose the prefix 'Al' for words starting with a solar consonant since due to assimilation with the following consonant, deletion of the prefix was one of the most frequent errors. Comparable recognition performance was achieved using word-based and morphologically decomposed language models, and since the errors made by the systems are different, combining the two gave a performance gain.

Index terms: Morphological decomposition, Arabic speech recognition

2. INTRODUCTION

As for other morphologically-rich languages such as Estonian, Finnish, German, Korean and Turkish [7, 21, 3], one of challenges of Arabic speech recognition is to deal with the huge lexical variety. For Arabic the combination of compounding, agglutination and inflection generate a large number of surface forms for a given root form. Morphological decomposition [13, 20, 22] has been proposed to address this problem increasing lexical coverage, thereby reducing errors that are due to words that are unknown to the system.

Generally speaking, a recognition vocabulary is simply a list of words as found in texts of the language. This view is a bit simplistic as it assumes that the texts have already been normalized, which in turn entails a variety of more or less important decisions [2, 4]. For morphologically rich languages there has been growing interest in using sub-word units to reduce the needed vocabulary size for a given lexical coverage. There are two main approaches to morphological decomposition, those based on the use of explicit linguistic knowledge and rules (for example, [18, 20, 22]), and unsupervised methods (for example, [12, 11, 3, 9]). Since the Arabic language has a relatively limited number of affixes, and rules can capture the manner in which they are applied, in this work the rules as implemented in the Buckwalter morphological analyzer are used [6, 10].

In the next section the variant methods for morphological decomposition are described, followed by a description of the audio and text training corpora used in the recognition experiments.

3. METHODOLOGY

Three variant methods for morphological decomposition were investigated. For all three the basis for decomposition is derived from the results of the Buckwalter morphological analysis [6]. In Buckwalter, the following affixes are decomposed (the Buckwalter transliteration codes are used here):

- 12 prefixes with 'Al': Al wAl fAl bAl wbAl fbAl ll wll fll kAl wkAl fkAl
- 11 prefixes without 'Al': w f b wb fb l wl fl k wk fk
- 6 negation prefixes: mA wmA fmA lA wlA flA
- 3 prefixes future tense: s ws fs
- suffixes (possessive pronouns): y, ny, nA, h, hm, hmA, hn, k, kmA, km, kn

In total there are 32 prefixes, 6 for negation, 3 for the future formed and 12 formed with the definite article, and 11 others without 'Al'. The suffixes in Arabic are personal pronouns, the objective form serves as a direct object of a verb, and as the possessive form serves as the complement of a noun.

In the first variant, a set of decomposition rules were applied to all words in the training texts that were identifiable by the Buckwalter morphological analyzer. Of the

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

Accepted after peer review of full paper

1137k distinct words in the training texts, 880K can be decomposed with the rules. About half of the remaining words are simple words, and the remainder have several possible decompositions (29%) or have a root that is not in the recognition dictionary (12%). Decomposition of a 200K lexicon results in a lexicon with 79K entries and reduces the out-of-vocabulary rate from 4.4% to 2%. If the decomposition rules are applied to the entire 1.1 M words, it is reduced to 270k forms (stems, affixes and uncomposed words). During decomposition, each affix that is split from the word root is marked by adding a "+" (to the end of prefixes and the start of suffices) to signify that it should be recomposed with the following or preceding word in the recognizer hypothesis.

In the first version (v1) the decomposition was applied to a list of 1.1 M words that were recognized by Buckwalter. Of these 880k were decomposed, and 256k remained unchanged. After decomposition, the word list was reduced to 270k forms (stems, affixes and uncomposed words). Following what has been done by others, in the second version (v2) the most frequent 65k words were never decomposed. This had the effect of blocking the decomposition of 35k words, which when added to the word list increased its size to 300k words.

In the third version (v3), on top of v2, the prefix 'Al' is not decomposed if the word begins with a solar consonant (the solar consonants in the Buckwalter code are: t, v, d, g, r, z, s, $, S, D, T, Z, l, n.). The reason to forbid the decomposition of 'Al' preceding words starting with a solar consonant is because the 'l' is assimilated with the following consonant and it is difficult to isolate a portion of the signal that clearly corresponds to the 'Al'. This problem is illustrated in the spectrogram in Figure 1 which is the excerpt (kaAn)ati AlT∼aA}irap Al$∼ir(aAEiyap) in Buckwalter code. The letters in parenthesis at the start and end provide the context. For the portion of interest ati AlT∼aA}irap Al$∼ir the first i was underlyingly a 'sukoun' (a mark which inhibits the pronunciation of a vowel). However, preceding the Al it is realized as an i (which is reduced to more or less a schwa) and the Al causes the following consonant to be realized as a geminate TT. This example shows a second gemination SS corresponding to the second Al. These type of phenomena are extremely difficult to model when the Al is allowed to be decomposed from the word, and explains why the Al was involved in so many of the errors in the first version. This restriction blocks decomposition of the prefix 'Al' preceding a solar letter if it is a simple prefix. If the prefix 'Al' is preceded by other prefixes, the other prefixes are split off and the 'Al' is kept with the stem.

For example, the original decomposition rules split the word wbAlslAm which has 3 prefixes w+b+Al+slAm into wbAl+ slAm, whereas the version 3 decomposition gives wb+ AlslAm.

We noticed that some of the words that were not able to be analyzed with Buckwalter were in dialectal Arabic. Adding seven dialect prefixes to the Buckwalter prefix table allows over 85% of these words to be decomposed.

In addition to processing the training texts and constructing language models, in order to build a complete system using the morphological decomposition, the affixes needed to be added to the pronunciation dictionary, and acoustic models trained with the decomposed lexical units. The pronunciation lexicon was extended to include all possible pronunciations of the affixes. One particular problem is handling the article 'Al' when it is followed by a solar consonant, since in this case the 'l' assimilated with the consonant. This phenomenon is taken into account within words by assigning a gemination mark to the consonant. To represent this in the decomposed prefix 'Al', contextual pronunciations are included for all solar consonants. For the acoustic models, the decomposition rules were applied to the transcripts of the audio training data, and several iterations segmentation and model estimation were carried out.

4. EXPERIMENTAL CONDITIONS

The training and test data are all from the Gale program, and distributed by LDC [1]. The audio training data used in this work are comprised of 1200 hours of manually transcribed broadcast data (1200h train). Roughly 60% of the data are classed as broadcast news (BN), that is typically well-prepared speech from announcers and reporters in speaking Modern Standard Arabic, and 40% is classified as broadcast conversation (BC), which tends to be more casual in style and has a higher proproation of dialectal Arabic.

The texts used for language model training are obtained from written sources and transcriptions of audio data. The written texts comprise more that 1.1 billion words from a variety of news sources, predominantly newspapers and newswire in Arabic. The transcriptions of audio data contain over 11 M words: 6.3 M words from BN and 4.8 M words of BC, and an additional 3.8 M words of Web transcripts of Aljazeera BC data.

The baseline recognition lexicon has 200k non-vocalized entries, each of which is associated with multiple vocalized forms, which in turn are associated with one or more phone pronunciations [15]. The pronunciations make use of 71 symbols, including 31 simple consonants, 30 geminate consonants, 3 long and 3 short vowels, plus 3 pseudo phones for non-linguistic events (breath, filler, silence). There are on average 8.6 pronunciations/word.

Results are reported on the Gale development and evaluation data sets from 2006 and 2007 (bnat06, bnad06, bcat06, bcad06, eval06, dev07, eval07). Each set contains 2 to 3 hours of audio data.

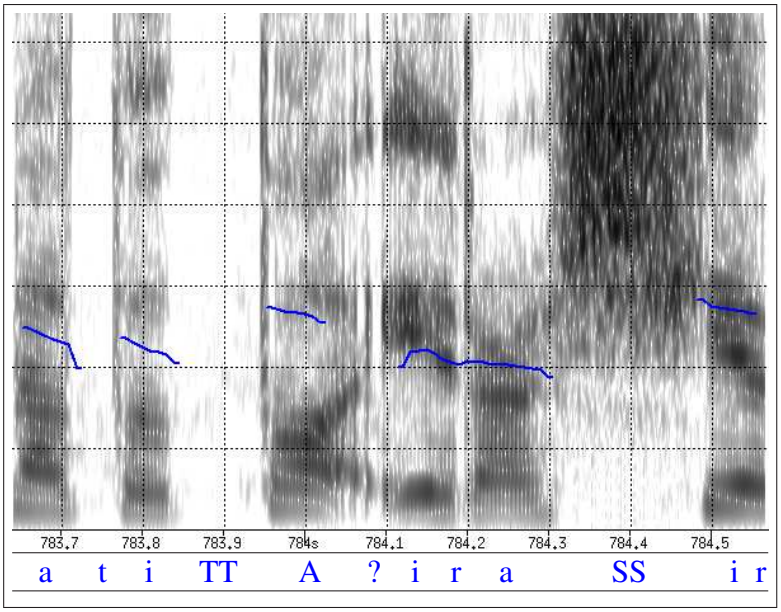

Figure 1: Example of assimilation of 'Al' preceding a solar consonant. The segment corresponds to the sequence (kaAn)ati AlT∼aA}irap Al$∼ir(aAEiyap) in Buckwalter code and is taken from the segment `LBC_NEWS_ARB_20060601_195801-0783.64-784.57`.

5. EXPERIMENTAL RESULTS

Initial experiments were carried out using the Jun'07 acoustic and language model training data. Acoustic models were trained on about 325 hours of manually transcribed data and 1000 hours with automatic transcripts. The language models were trained on almost 1 B words of texts, including 2 M words of audio transcripts.

The three versions of decomposition were applied to the training transcripts, and three sets of word-position dependent acoustic models were estimated, specific to each versions. The WER of the reference word based system with MLE training was 22%. With the first decomposition method that simply splits all affixes, the WER is increased by 2% absolute. By forbidding the decomposition of the most frequent 65k words (v2) most of these errors are avoided, as shown in the entry 'Decomposition version 2, LM' which used acoustic models trained for the word based system. The second entry for version 2 shows the effect of retraining acoustic models with decomposed transcripts. Applying the 3rd version of decomposition rules prevents the decomposition of 11k solar words. After applying these to the full 1.1 M word list, the recognition vocabulary contains 320k entries (stems, affixes and uncomposed words). This decomposition was applied to the training texts and language models were built. The WER with the LM built with the version 3 decomposition is given in the bottom of Table 1. The WER is reduced by 0.7% compared to the v2 decomposition (LM only).

In a next set of experiments, the acoustic model training data was updated to make use of additional manually transcribed data distributed by LDC, that is about about 1200 hours of speech. Recognition results of a single decoding pass with unsupervised acoustic model adaptation are given in Table 1 for the Gale bnat06 dataset. The WER of the reference word based system under the same condition is 20.9%. The versions 1 and version 2 decompositions are both seen to still degrade performance, but there is a small gain with the 3rd method that does not decompose the prefix 'Al' preceding solar consonants. This method has been adopted for the remainder of the experiments reported here.

Next, making use of additional language modeling

Condition	*Vocabulary size*	*WER*
Reference word based	200k	22.0
Decomposition v1	270k	24.0
Decomposition v2, LM	300k	22.3
Decomposition v2, LM + AM	300k	22.1
Decomposition v3, LM	320k	21.6

Table 1: WER on the bnat06 dataset with the reference 200k word based system (325h train), the system with morphological decomposition versions 1, 2 & 3 for a single decoding pass with acoustic model adaptation.

Condition	*WER*	*Vocabulary size*
Reference word based	20.9	200k
Decomposition v1	23.7	270k, baseline
Decomposition v2	21.8	300k
Decomposition v3	20.5	320k

Table 2: WER on the Gale bnat06 dataset with the reference 200k word based system, the system with three morphological decomposition versions for a single decoding pass with acoustic model adaptation (1200h train).

Conditions	*bnat06*	*bnad06*	*bcat06*	*bcad06*	*eval06*	*dev07*	*eval07*
Baseline	16.7	15.5	22.8	20.4	19.3	12.4	13.7
Decomposition	16.7	15.3	23.1	20.6	19.4	12.2	13.8
Combination	16.1	14.9	22.3	19.7	18.5	11.8	13.2

Table 3: Word error rates for the 290k word based LM (baseline) and the 290k LM with morphological decomposition for different data sets. All conditions use MMIE trained acoustic models and a NN language model.

data, an updated 290k recognition word list was selected and optimized on the combined development data from 2006 and 2007. Applying the v3 decomposition rules to the 290k test lexicon, reduces it to 173k entries. The word list was then extended to 290k entries (LM290kmd) by adding the forms produced by decomposing the most probable words recognized by Buckwalter not already covered.

Table 3 reports results using MMI trained acoustic models (on the 1200 hours of manually transcribed data), developed for the word-based system, that is the training transcriptions use a word representation. Results are given for all Gale development sets with neural net language models that has been estimated on the texts that have been morphologically decomposed and for the baseline word based NN LM [19]. From these results it can be seen that the two language models give quite comparable results. The results obtained by combining the two the models using Rover are given in the third row of this table. Compared to the baseline system the average word error reduction across all test sets is about 0.6%.

6. SUMMARY

Although there has been growing interest in using some form of morphological decomposition to address speech recognition challenges for morphologically rich languages, our experience is that brute force methods may reduce performance by creating more acoustic confusability of small units [17] and by their inability to capture some coarticulation phenomena. For Arabic, the realization of the definite article 'Al' is particularly susceptible to such coarticulation effects, for which it was found to be effective to inhibit decomposition prior to solar consonants. Since the word based and decomposed transcription systems do not make the same errors, combining the two reduced the word error rate of a state-of-the-art Arabic STT system.

REFERENCES

[1] http://projects.ldc.upenn.edu/gale/index.html

[2] G. Adda, M. Adda-Decker, J.L. Gauvain, & L. Lamel, "Text normalization and speech recognition in French," *EuroSpeech'97*, **5**:2711-1714, Rhodes, 1997.

[3] M. Adda-Decker, "A corpus-based decompounding algorithm for German lexical modeling in LVCSR," *ISCA Eurospeech'03*, Geneva, September 2003.

[4] M. Adda-Decker, L. Lamel, "The use of lexica in automatic speech recognition," chapter in *Lexicon Development for Speech and Language Processing*, 235-266, F. Van Eynde et D. Gibbon, eds., Pays Bas: Kluwer, 2000.

[5] E. Arisoy, H. Sak, & M. Saraclar, "Language Modeling for Automatic Turkish Broadcast News Transcription," *Interspeech'07*, Antwerp, 2007.

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

[7] K. Carki, P. Geutner & T. Schultz, "Turkish LVCSR: towards better speech recognition for agglutinative languages," *ICASSP'00*, 3688-3691, Istanbul, 2000.

[8] M. Creutz et al., "Analysis of Morph-Based Speech Recognition and the Modeling of Out-of-Vocabulary Words Across Languages," *NAACL-HLT*, 380-387, 2007.

[9] M. Creutz & K. Lagus, "Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora using Morfessor 1.0," *Computer and Information Science*, Report A81, 27, 2005.

[10] A. Ghaoui, F. Yvon, C. Mokbel, G. Chollet, "On the use of morphological constraints in n-gram statistical language model," *Interspeech-2005, 1281-1* 2005.

[11] J. Goldsmith, "Unsupervised learning of the morphology of a natural language," *Computational linguistics*, **27**(2):153-198, 2001.

[12] Z.S. Harris, "From Phoneme to Morpheme, " *Language*, **31**:190-222, 1955.

[13] K. Kirchhoff et al., "Novel approaches to Arabic speech recognition - final report from the JHU summer workshop 2002," John-Hopkins University, Technical Report, 2002.

[14] K. Kirchhoff & R. Sarikaya, "Processing morphologically rich languages," *Interspeech 2007 workshop*, Antwerp.

[15] A. Messaoudi, J.L. Gauvain & L. Lamel, "Arabic Broadcast News Transcription using a One Million Word Vocalized Vocabulary," *ICASSP'06*, Toulouse, May 2006.

[16] L. Lamel, A. Messaoudi & J.L. Gauvain, "Improved Acoustic Modeling for Transcribing Arabic Broadcast Data," *Interspeech'07*, 2077-2080.

[17] T. Pellegrini & L. Lamel, "Using phonetic features in unsupervised word decompounding for ASR with application to a less-represented language," *InterSpeech'07*, Antwerp, August 2007.

[18] H. Schmid, "Probabilistic Part–of–Speech Tagging Using Decision Trees," *Conference on New Methods in Language Processing*, Manchester, 1994.

[19] H. Schwenk, "Continuous space language models," *Computer Speech and Language*, 21:492-518, 2007.

[20] D. Vergyri, K. Kirchhoff, K. Duh and A. Stolcke, "Morphology-based language modeling for Arabic speech recognition," *Proc. ICSLP*, 1252-1255, Jeju, 2004.

[21] E. Whittaker & P. Woodland, "Particle-based language modelling," *ICSLP'00*, Beijing, 2000.

[22] B. Xiang, K. Nguyen, L. Nguyen, R. Schwartz & J. Makhoul "Morphological Decomposition for Arabic Broadcast News Transcription" *Proceedings of ICASSP*, **I:**1089-1092, Toulouse, May 2006.