Investigating Automatic Decomposition for ASR in Less Represented Languages

Thomas Pellegrini and Lori Lamel

LIMSI-CNRS, BP133
91403 Orsay cedex, FRANCE
{thomas.pellegrini, lamel}@limsi.fr

Abstract

This paper addresses the use of an automatic decomposition method to reduce lexical variety and thereby improve speech recognition of less well-represented languages. The Amharic language has been selected for these experiments since only a small quantity of resources are available compared to well-covered languages. Inspired by the Harris algorithm, the method automatically generates plausible affixes, that combined with decompounding can reduce the size of the lexicon and the OOV rate. Recognition experiments are carried out for four different configurations (full-word and decompounded) and using supervised training with a corpus containing only two hours of manually transcribed data.

Index Terms: less represented languages, speech recognition, Amharic, lexicon construction, word decomposition.

10.21437/Interspeech.2006-89

1. Introduction

With today's technologies, large corpora of transcribed speech and large amounts of textual data are still required to build automatic speech recognition (ASR) systems. While the Internet is a great source of raw textual and audio data that can be exploited for building acoustic and language models using semi-automatic methods, a great majority of the world's languages suffer from poor representation on the Internet. The Amharic language is an example of a less represented language, for which only small quantities of written texts are available. This article reconsiders what units should be represented in the recognition system, i.e. what is the definition of a word. A corpus-based method is described and used to help select recognition units for the Amharic language. A set of affixes is automatically selected by measuring the number of new words introduced in the lexicon when words are decompounded. The automatic selection of affixes is not straightforward. Different selection criteria are discussed and the impact of decomposition on lexical coverage is measured. The generalizability of the method is demonstrated by applying the affix selection process to the Arabic language, which belongs to the same Semitic family of languages.

In the next section, the audio and textual resources used in this work are described. This is followed by a description of the affix selection method along with corpus based studies. Automatic speech recognition experiments are presented using the different representations in Section 4, with a training corpus of only two hours of manually transcribed data.

2. Audio and textual data

The Amharic language was used in the study, being representative of languages for which at this time only limited resources are available. Amharic is the official language of Ethiopia and has about 14 million speakers (source: omniglot.com). Although it is a semitic language like Hebrew and Arabic, its writing, which developed from the Ethiopian classical language Ge'ez, is a syllabic left-to-right script. Amharic has 34 basic symbols, for which there are 7 vocalizations: /ɛ/, /u/, /i/, /a/, /e/, /ə/ and /o/, referred to as the seven orders. The basic symbols are modified in a number of different ways to indicate the different vocalizations. 85% of the syllables represent a CV sequence (C for consonant and V for vowel), one symbol represents the complex sound /ts/V and the reminder represent CwV sequences (where w is a semi-consonant). There are various recent studies on speech recognition and speech processing for Amharic [1, 2, 3], a new resource web portal for Amharic corpora has also been created. [1]

Table 1: *Characteristics of the audio corpus (number of hours, number of speakers, and number of words for each audio source.*

Source	*Training*	*Development*
Deutsche welle	24h 6mn	1h 20mn
Radio Medhin	11h 8mn	37mn
# speakers	200	15
# words	233k	14k

Compared to other languages for which models and systems have been developed [4], the Amharic audio corpus is quite small, containing a total of 247k words with 50k distinct lexemes. It is comprised of 37 hours of broadcast news data from Radio Deutsche Welle (25h) and Radio Medhin (12h), recorded during the period from January 2003 through February 2004. The shows have been transcribed by native Ethiopian speakers. Two hours of data taken from the latest shows for each audio source were reserved for development test. Table 1 summarizes the characteristics of the audio corpus in terms of the number of hours by source, the number of distinct speakers, and the number of words for both the training and the development subsets. More hours were recorded of Deutsche Welle since it has a greater diversity of speakers than does Radio Medhin.

In addition to the transcriptions of the audio data, about 4.6 Millions of words from the newspaper and web texts dating from 1996 to 2005 from three sources have been used for language model training. These texts include about 1.72M words the newspaper Ethiozena archives (1996-2004), 1.1M words of Deutsche Welle web texts (2003-2004) and the newspaper Ethiopian Reporter (1.8M words recent dating from 2004-2005). The text sources were used to select the recognition vocabulary and to train language models, as is further described in Section 4.

[1] http://corpora.amharic.org/

Table 2: *Statistics for two potential suffixes: number of occurrences as a word, as an affix, number and percentage of roots not in the initial lexicon.*

	# Occurrences		New roots	
Affix	*as a word*	*as an affix*	#	%
+woCx	92	49214	1105	2.2
+wana	4108	725	282	38.9

Table 3: *Number of affixes proposed by the Harris algorithm, and after the first and second selection step.*

# Affixes	*Initial*	*After step 1*	*After step 2*
Prefixes	500	495	175
Suffixes	386	371	126
Total	886	866	301

3. Word decompounding

An initial 114k word-based lexicon was selected, comprised of all distinct words in a two-hour subset of the training data transcriptions and all the words occurring at least three times in the newspaper and web texts. The out-of-vocabulary (OOV) rate of the development corpus with this word list is 18%, which is very high. There are a total of over 340k distinct words in the texts. The high OOV rate even with a large word list led to an investigation of how effective word decompounding of affixes could be for the Amharic language. Studies on word decompounding have been described for languages in which the word compounding generation process is frequent, such as for German [5]. Compounding in Amharic results from the addition of prefixes and suffixes that are grammatical morphemes like pronouns, possessive and demonstrative adjectives, prepositions and postpositions.

3.1. Method for affix selection

The automatic detection of affixes is interesting since no linguistic information specific to the target language is required, thereby the method can be used for essentially any language with little adaptation. The Harris algorithm [6] originally applied to phoneme strings, can also be used to detect morpheme boundaries from a simple list of words in the target language. It relies on the universal property that the number of potential distinct letters which can follow any given word beginning reduces rapidly with the length of the word-initial substring. Generally speaking, as the number of letters in the substring increases the number of possible successors decreases. If the number of successors increases for a particular substring, then the substring is a potential prefix that can be recombined with other morphemes that begin with various distinct characters. The algorithm counts the number of distinct successor characters for each prefix of a given length and proposes morpheme boundaries when a local maximum is found. The goal is not to carry out a morphological analysis, but to determine a list of potential affixes. Decompounding the words with the most frequent affixes allows the number of distinct lexemes to be reduced while increasing the representation of some infrequently observed n-grams [5].

3.2. Application to Amharic

The algorithm was applied to the 114k words in the Amharic lexicon. 500 prefixes and 386 suffixes were found. Affixes were selected in two steps. First, for affixes that occur both as affixes and as words, the number of occurrences in the text corpus as a word were compared to the number of occurrences as an affix. Affixes occurring more often as a word than as an affix were excluded. Second, the utility of splitting the affix from the rest of the word was considered by counting the percentage of times that decompounding creates new roots (i.e. creates words that were not in the original word list). Table 2 shows two examples of potential suffixes and their counts. The '+' sign has been added to the affixes in order to ensure the possibility of recombining affixes and roots back into entire words. The first affix '+woCx' occurs almost 50k times as a suffix and leads to only 2.2% of new lexemes when decompounded in the lexicon. The second potential suffix appears much more frequently as a word (4.1k times) than as a suffix (725 times), so this one is rejected. Of the almost 900 affixes proposed by the algorithm, only 20 were excluded after the first step since they occurred more frequently as words than as affixes. At the second step, it was decided that in order to achieve the goal of reducing the size of the word list it only makes sense to split affixes if they generate fewer new lexemes than are removed by decompounding. Therefore in these experiments only affixes that generate less than 50% of new lexemes were selected, where the lexeme counts are weighted by their frequency in the texts. Table 3 gives an overview of the number of affixes at each step of the selection process.

The 'split_all' curve in Figure 1 shows the number of new lexemes created by the decompounding as a function of the number of prefixes used, ranked by decreasing frequency in the texts. The curves for the suffixes are very similar and are therefore not presented here. The most frequent prefixes create many new lexemes. Two criteria for decompounding are shown. New lexemes

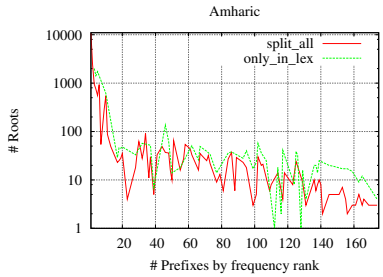

Figure 1: *Number of roots as a function of the number of prefixes ranked by frequency. For the 'split_all' curve new lexemes are incrementally added to the word list. For the 'split_only_in_lex' curve only words that generate lexemes already in the word list are split.*

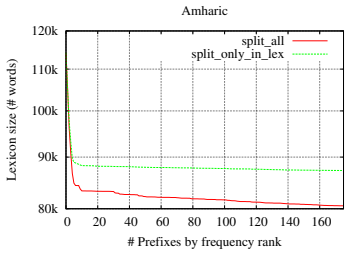

Figure 2: *Lexicon size as a function of the number of decompounded prefixes (ranked by frequency). For the 'split_all' condition, new roots are added to the lexicon before decompounding the next prefix. The initial lexicon has 114k entries.*

Table 4: *Impact of decompounding on lexicon size and OOV rate.*

	# Affixes	*Lexicon size*	*OOV(%)*
Entire words	0	114.2k	18.0%
Prefixes	175	89.5k	15.1%
Suffixes	126	91.5k	14.1%
Affixes	301	68.1k	12.1%

can be incrementally added as they are created ('split_all' curve) or words can be split only when they generate lexemes that are already in the word list ('split_only_in_lex' curve). It can be seen that in both cases the most frequent prefixes generate the most roots, and even in the 'split_all' case the number of new lexemes generated by the decomposition decreases rapidly. For the least frequent fewer than 10 new roots are created. For the Amharic language where the grapheme-to-phoneme conversion is straightforward [7] adding new lexemes to the word list is not problematic.

Figure 2 shows the impact of word decompounding with the selected affixes on the lexicon size. The prefixes are ordered by their frequency in the texts and are split one at a time. For the curve 'split_all', newly created roots are added to the word list and the original lexemes are removed. The most frequent prefixes are short, single syllable prefixes and systematically splitting them has a large impact on the word list size. As the frequency rank of the prefix decreases, the effect of splitting is smaller. For the curve 'split_only_in_lex', only those words which when decompounded have the root in the word list are split. As a results, the number of lexical entries is seen to asymptote more quickly. Table 4 shows the impact of word decompounding with the selected affixes on the lexicon size and the OOV rate, when splitting affixes only for remaining roots of two syllables at least (in figure 2, no condition on roots was applied). With all 301 affixes, the number of distinct lexemes decreases from 114k to 68.1k words. The OOV rate is reduced by over 30% from 18.0% to 12.1%.

3.3. Application to Arabic

In order to validate the basic methodology, a similar study was carried out decompounding prefixes for the Arabic language. Using an initial lexicon with 101k entries, a smaller number of proposed prefixes are found than were proposed for Amharic. These correspond to Arabic prefixes (Al, wa, bi, li, waAl, ka, lil, sa, fa,...). Figures 3 and 4 show respectively the number of roots after splitting and the impact of word decompounding with the selected affixes on the lexicon size. As for Amharic, for the 'split_all' curve new lexemes are incrementally added to the word list and for the 'split_only_in_lex' curve, only words that generate lexemes already in the word list are split. The smaller reduction in lexicon size for Arabic can be attributed to the significantly smaller number of selected prefixes.

4. Recognition experiments in Amharic

Speech recognition experiments were carried out to compare a standard full-word representation with three different word representations using word decompounding. The speech recognizers all have two decoding passes with an unsupervised adaptation of the acoustic models after the first pass [8].

The language models are Kneyser-Ney smoothed trigram models, and result from the interpolation of two component LMs: one estimated on the web texts and the other on the manual transcripts of the audio data. The interpolation coefficient was optimized by measuring the perplexity of the LM on the dev test transcripts.

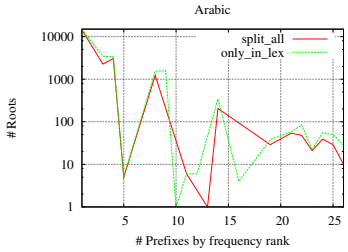

Figure 3: *Number of roots as a function of the number of prefixes ranked by frequency in the lexicon.*

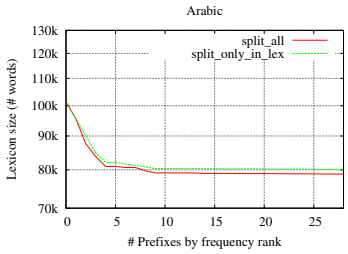

Figure 4: *Lexicon size as a function of the number of decompounded prefixes ranked by frequency in the lexicon. The initial lexicon has 101.3k entries*

4.1. Influence of training corpus size

Figure 5 shows the word error rate (WER) as a function of different quantities of manually transcribed data used to train the acoustic models. The WER is measured on the dev corpus (2h of speech, 14.1k words). For each point on the curve, the language model and the word lexicon were re-estimated using only the transcriptions used to train the acoustic models. There is a rapid decrease in WER from 10 minutes to 2 hours of training data, from over 76% to about 50.0%. In the range above 2 hours of data, the WER decreases more slowly to a WER of 41.5% with 10 hours of data. The abrupt change in the WER from 1h to 2h of training data is due to the acoustic models, and not the LM since the same WER is obtained when the LM corresponding to 1h of speech is used. This may be due to a better representativity of the speakers in the larger audio set. Considering this, further experiments using affix decompounding combined with supervised training were carried out with the same two hours of manually transcribed training data with a total of 16.6k words in transcripts.

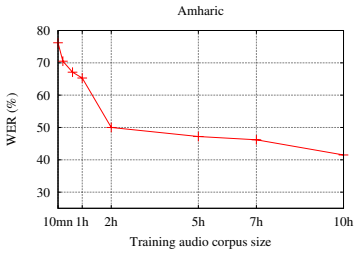

Figure 5: *Word Error Rate vs quantity of training transcriptions*

Table 5: *Word error rates before (WER+) and after (WER) recombination.*

Representation	WER+ (%)	WER (%)
full words	-	50.0
prefixes	38.7	**44.8**
suffixes	43.1	46.5
affixes	39.8	47.8

Table 6: *WER, Insertion, Deletion and Substitution rates before (top) and after word recombination (bottom).*

Representation	WER+ (%)	I+	D+	S+
prefixes	38.7	3.0	8.3	27.5
suffixes	43.1	3.6	10.5	29.0
affixes	39.8	3.7	10.0	26.1

Representation	WER (%)	I	E	S
full words	50.0	1.3	10.9	37.8
prefixes	44.8	1.8	8.2	34.8
suffixes	46.5	1.7	8.6	36.2
affixes	47.8	1.8	8.9	37.1

4.2. Experiments with word decompounding

Three representations involving word decompounding were investigated: with only the 175 selected prefixes separated; with only the 126 suffixes separated; and both the prefixes and suffixes separated (301 affixes in total). Decompounding words into smaller units changes the word boundaries. Since the acoustic models are word-position dependent, new acoustic models were trained for each representation. These position-dependent triphone models cover about 3.1k contexts, with 8 Gaussians per state. Language models for each condition were also estimated. Table 5 shows the WER for the four systems, the first column (WER+) gives the WER before recombining the morphemes into words and the second column gives the WER after recombination. Looking at the outputs of the systems, the affixes seem to be well recognized, because of their small length and their high frequency. Recombining words increases the WER showing that errors are mainly on the roots. When recombining, the number of words in the sentences decreases so that the number of errors is shared by less units. Nevertheless gains after recombination are observed in the WER compared to the full word representation (50% WER). The best WER was obtained with the representation splitting prefixes which had an absolute gain of 5.2%. Combining the separation of both prefixes and suffixes gave the smallest gain (2.2%). Using both prefixes and suffixes reduces the language model context and therefore may confuse the system. Longer spam n-grams may help to address this problem.

Table 6 gives the repartition of the errors between the Insertion, Deletion and Substitution rates, respectively before and after word recombination. With the affixes separated, more insertions and deletions are observed since the texts contains many more little words than the initial texts with a full words representation. The main gains are obtained on the substitution rates.

Table 7 shows an example sentence that was correctly recognized with the separated prefix system but the full-word system hypothesized two words instead of three. The reference sentence, shown in bold, is composed of three words for the full-word representation (S_{word}) and four words when the prefixes are separated (S_{prefix}). The figure gives the log-likelihood (llh) of each word sequence with the corresponding LMs. For the word based system, the correct sentence (reference) has a less good likelihood than the hypothesis. For the separated prefix system, the words 'lajx' and 'jE+' are among the most frequent words in the texts. The second best hypothesis with the first two words glued together '?iraKxlajx' is much less frequent and has a much lower likelihood. For the full-word system the high unigram probability of the word 'lajx' is not enough high to output the correct sentence. Since the word 'jESxgxgxrx' is rare in the texts, the separation of the prefix 'jE+' is advantageous.

Table 7: *Example of a sentence recognized correctly by S_{prefix} and incorrectly by S_{word}. The correct and incorrect word strings are given along with their log-likelihoods.*

System	*Sentence*			*llh*
S_{word}	?iraKxlajx		jESxgxgxrx	-9.5551
	?iraKx	**lajx**	**jESxgxgxrx**	-9.6559
S_{prefix}	**?iraKx**	**lajx**	**jE+ Sxgxgxrx**	-9.2613
	?iraKxlajx		jE+ Sxgxgxrx	-10.1367

5. Summary

This paper has described the use of an automatic decomposition method based on the Harris algorithm that is seen to automatically generate plausible affixes, which when combined with decompounding can reduce the size of the lexicon and the OOV rate. Corpus-based studies were carried out for the Amharic language selected as being representative of languages for which only a small quantity of linguistic resources are available. Recognition experiments were reported for four different configurations (full-word and decompounded) using a training corpus containing only two hours of manually transcribed data. The system with prefix decomposition was shown to reduce the relative word error rate after recombination by about 10% compared to the full word models.

6. References

[1] W. Menzel S.T. Abate and B. Tafila, "An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition," in *Interspeech*, Lisboa, 2005.

[2] B. Gamback H. Seid, "A speaker independent continuous speech recognizer for Amharic," in *Interspeech*, Lisboa, 2005.

[3] S. Eyassu and B.Gamback, "Classifying Amharic news texts using self-organizing Maps," in *ACL05 Workshop on computational Approaches to Semitic Languages*, Ann Arbor, 2005.

[4] R. Schwartz and al., "Speech recognition in multiple languages and domains: The 2003 bbn/limsi ears system," in *ICASSP*, May 2004.

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

[6] Z. Harris, "From Phoneme to Morpheme," in *Language 31*, 1996, pp. 190–222.

[7] T. Pellegrini and L. Lamel, "Experimental detection of vowel pronunciation variants in Amharic," in *LREC*, Genoa, 2006.

[8] J.L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, vol. 37, pp. 89–108, 2002.