Using phonetic features in unsupervised word decompounding for ASR with application to a less-represented language

Thomas Pellegrini, Lori Lamel

LIMSI-CNRS, BP133, 91403 Orsay cedex, FRANCE

thomas.pellegrini@limsi.fr, lamel@limsi.fr

10.21437/Interspeech.2007-502

Abstract

In this paper, a data-driven word decompounding algorithm is described and applied to a broadcast news corpus in Amharic. The baseline algorithm has been enhanced in order to address the problem of increased phonetic confusability arising from word decompounding by incorporating phonetic properties and some constraints on recognition units derived from prior forced alignment experiments. Speech recognition experiments have been carried out to validate the approach. Out of vocabulary (OOV) words rates can be reduced by 30% to 40% and an absolute Word Error Rate (WER) reduction of 0.4% has been achieved. The algorithm is relatively language independent and requires minimal adaptation to be applied to other languages.

Index Terms: automatic speech recognition, unsupervised word decompounding, less-represented languages

1. Introduction

In some languages it is common to generate words by the compounding of smaller units that are primarily lexical morphemes (such as in German or Turkish) or mostly grammatical morphemes (for example, Semitic languages such as Arabic or Amharic). For automatic speech recognition (ASR), word compounding poses at least two problems. The first concerns lexical coverage, since high out-of-vocabulary (OOV) rates are obtained even with quite large vocabularies (containing over 65k words). The second concerns language modeling, where it can be difficult to have reliable n-gram estimates for infrequent words. To address these issues, word decomposition has been investigated in many studies and for various languages such as German [1], Turkish, Finnish and Estonian [2], and Dutch [3].

High OOV rates and poor language model estimation are also problems faced when developing technologies for less-represented languages for which little data in an electronic form are available. Most of the world's languages suffer from poor representation on the web, which is being used more and more as the primary source for collecting data (principally texts) for building ASR systems. This study reports on experiments carried out with the Amharic language, the official language of Ethiopia, since it is both a less-represented language and a language in which grammatical compounding is frequent. In a previous study [4], improvements in WER were reported by decompounding words using Harris' algorithm [5] when building a system with very limited amounts of audio training data (2h, 17k tokens). Further experiments carried out using more training data (35 hours, 240k tokens) achieved worse performances when decompounding lexical units. This result has also been reported by others, for example in [6, 7] and one hypothesis is that phonetic confusability is increased when generating small recognition units, and that these units are frequent and therefore highly probable in the language model. The work described here reports on modifications made to the statistical word decomposition paradigm to address the problem of phonetic confusability.

The term "morph" is used in this article to name either words or word splits, however the splits are not always true morphemes in a linguistic sense.

2. Unsupervised word decompounding

Automatic word decompounding is investigated as a means to help select recognition units in an almost language-independent manner. This work is an extension of the Morfessor algorithm, as implemented in the open source Perl program called 'Morfessor 1.0' from the Helsinki University of Technology [8].

2.1. Baseline Morfessor 1.0 algorithm

An overview of the basics of this algorithm is provided here, for further information refer to [8]. The program has two purposes: first, the training of a word segmentation model given a lexicon with optional frequency counts. Training uses a maximum a posteriori (MAP) criterion based on several text properties. Second, a previously learnt decomposition model can be used to decompound a new word list. Words that are not in the model can also be decomposed since the algorithm is Viterbi-like. This search algorithm relies only on the morph frequencies.

During model training, the algorithm tries to iteratively maximize the following estimate:

$$\underset{\mathrm{L}}{\operatorname{argmax}}\ P(\mathrm{L}|\mathrm{corpus}) = \underset{\mathrm{L}}{\operatorname{argmax}}\ P(\mathrm{corpus}|\mathrm{L})P(\mathrm{L}) \quad (1)$$

where $P(\mathrm{corpus}|\mathrm{L})$ is the maximum likelihood estimate of the corpus given a lexicon L, based on the word frequencies. $P(\mathrm{L})$ is the *a priori* probability of the lexicon L, i.e., the probability of getting M distinct morphs $m_1, ..., m_M$. In the baseline algorithm, two properties are used to estimate $P(\mathrm{L})$, these are the word frequency f_m and character sequence s_m probabilities as shown in equation 2. The modifications affect the properties used in equation 2.

$$P(\mathrm{L}) = P(f_{m_1}, ..., f_{m_M})P(s_{m_1}, ..., s_{m_M}) \quad (2)$$

2.2. End-of-word probability

In the baseline Morfessor program, the character probabilities $P(s_m)$ of equation 2 are static constants, calculated only once during model initialization, as the simple ratio of the number of occurrences of the character divided by the total number of characters in the corpus. These are independent of word position. To represent the word boundary, a space character is added to each lexical entry. The end-of-word probability is the probability of the space character, and has the same value for all words and morphs in the corpus.

Inspired by Harris' algorithm [5], we propose replacing this static probability by the probability P_H defined in equation 3. The probability that a word beginning WB is a morph is defined as the ratio of the number of distinct letters $L(\text{WB})$ which can follow WB over the total number of distinct letters L.

$$P_H(\text{WB}) = L(\text{WB})/L \tag{3}$$

This definition favors short morphs, which is potentially interesting for languages where the word compounding generation process results from the addition of prefixes and suffixes that are grammatical morphemes such as pronouns, possessive and demonstrative adjectives, prepositions and postpositions. For languages in which the compounding is mainly lexical, the morphs are generally longer. The original Harris' algorithm is more appropriate for such languages since it searches for a local maximum of $L(\text{WB})$ [1].

3. Modified algorithm for ASR

3.1. Distinctive feature motivated property

All the properties used in the Morfessor program are based on written language and do not incorporate any "oral" properties that could be useful for ASR. Given the confusions observed in prior studies [10], a phone-based feature was added to the $P(\text{L})$ term of equation 2. This property aims to represent the phonetic confusability between lexical units. It is theoretical and relies on some distinctive features (DF) of the phones used in the language of study. Equation 4 gives the definition for a morph m_k. As a first approach, we chose to limit the computation to the vowels and to the morphs that share the same consonantal root. The distance $D_{DF}(m_k)$ is the range $[0, 1]$.

$$D_{DF}(m_k) = \prod_{j=1}^{j=N_k-1} D_{DF}(m_k, m_j) \tag{4}$$

with

$$D_{DF}(m_k, m_j) = \prod_{l=1}^{l=V_k} \frac{\Delta_{kl,jl}}{C} \tag{5}$$

N_k is the number of morphs that share the same consonantal root, $\Delta_{kl,jl}$ is the number of different DFs in the l^{th} vowel of morphs m_k and m_j, and C is the total number of distinct DFs. To evaluate $\Delta_{kl,jl}$, one can use DF tables found in phonetics literature, for example in [9]. The distinctive features used in this study only concern vowel and are given in Table 2.

3.2. Phonetic confusion constraint

The DF property is theoretical and therefore does not account for the phonological variation observed in real world speech, such as in the choice of vowel alternatives. In [10] syllabotactic alignments were studied in order to determine the most frequent confusions at the syllable level. For each syllable, the vowel that was most often substituted by the aligner was determined. These confusion pairs provide an additional means of preventing phonetic confusion amongst units arising from the decompounding.

During the decompounding process, word splits that differ from other morphs by only one syllable are compared. If the pair of syllables is among the most frequently confused pairs found in the alignment study, the split is forbidden.

3.3. Summary of the different options

The different options investigated with the decompounding algorithm are summarized in Table 1. The three configurations M, M H, M H DF are compared both with and without the confusion constraint Cc.

Table 1: *Decomposition options compared in this study.*

Option	*Comment*
BL	Baseline word based system, no decompounding
M	Baseline Morfessor 1.0
M H	M + modified 'Harris'
M H DF	M H + distinctive features parameter
Cc	+ confusion constraint

4. Application to Amharic

The Amharic language is an example of a less-represented language, for which only small quantities of written texts are available. There are some recent studies on speech recognition and speech processing for Amharic [11, 4], and a web resource portal for Amharic corpora has also been created (http://corpora.amharic.org/). Grapheme-to-phoneme conversion is straightforward for this language. Amharic has a CV (consonant vowel) structure, with 85% of the syllables representing a CV sequence. One symbol represents the complex sound /ts/V and the reminder represent CwV sequences (where w is a semi-consonant). In this study, Cw has been considered as a single phone. Given the CV structure of the Amharic language, splits are allowed only after a vowel.

The distinctive features used with the 'DF' option are shown in Table 2. Based on vowel confusions reported in a previous study [10], in this study /a/ is considered non-tense.

4.1. Audio and textual data

Compared to other languages for which models and systems have been developed [12], 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. The data were transcribed by native Ethiopian speakers, and two hours of data taken from the latest shows were reserved for development test. Table 3 summarizes the characteristics of the audio corpus in terms of the number of hours by source, the number of distinct speakers, and the total number of words for both the training and the development subsets.

In addition to the transcriptions of the audio data, about 4.6 million words of newspaper and web texts have been used for language model training. Over 340k distinct words are found in these texts.

4.2. Decompounding the training texts

When building a recognition lexicon from training texts, we apply a frequency cut-off to get rid of misspelled words and artifacts. The basic idea in this study consists of building a decompounding model for a reference lexicon, and then using this

Table 2: *Distinctive features used with the algorithm.*

DF	*Vowels*						
IPA	ɛ / ə	u	i	a	e	ə/ i	ɔ
high	0	1	1	0	0	0	0
front	1	0	1	0	1	0	0
back	0	1	0	1	0	0	1
round	0	1	0	0	0	0	1
tense	0	1	1	0	1	0	1
reduced	0	0	0	0	0	1	0
long	0	1	1	1	1	0	1

Table 3: *Characteristics of the audio corpus (number of hours, speakers, and total 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

model to decompose all words in the corpus without any frequency cut-off. A new reference lexicon is then built applying a frequency cut-off. The OOV rate may decrease since OOV words may have been decompounded. The number of lexical tokens in the training text corpus is also increased with this method.

An initial 133k word-based lexicon was selected, comprised of the 50k distinct words in the training data transcriptions and all words occurring at least three times in the newspaper and web texts. The out-of-vocabulary rate of the development corpus with this word list is 7%, which is quite high compared to the OOV rates obtained for well-represented languages which are typically around 1-2%.

Table 4 shows the number of morph types for the different decompounding options listed in Table 1. Since a morph may exist both as a word and as an affix, the explicit use of this information is investigated by adding a '+' sign to prefixes found by the algorithm in order to ensure the possibility of recombining morphs back into entire words. Since the word and affix entries corresponding to the morph will have the same pronunciations in the recognition lexicon, the choice between forms is made by the language model. The third column gives the number of types when no explicit distinction is made between words and affixes (i.e., no '+' sign is added during decomposition). The difference between the second and the third columns is the number of morphs that also are words. The Harris option (H) gives the smallest lexicon with about 90k units. When prefixes are explicitly marked (as shown in the second column, '+'), the reduction in lexical size is only about 5% with the confusion constraint (Cc) and 30% without it, illustrating that most affixes are also words.

Table 4: *Number of morph types in the lexicons with and without '+' for different decompounding options.*

Options	# Morph Types '+'	# Morph Types
BL	0	133384
M	95937	70267
M Cc	128239	109694
M H	90740	65421
M H Cc	126105	107123
M H DF	94198	69038
M H DF Cc	128404	110320

Figure 1 shows the number of types (top) and tokens (bottom) as a function of their length in phones, for the different decompounding options. The BL curve (in black) is the baseline curve, where words are not decompounded. The other curves, for which words were decompounded and prefixes marked with a '+' sign, show a noticeable shift to smaller word lengths. The curves with and without the 'Cc' option form two distinct groups. The 'Cc' curves (drawn with 'x') have substantially more distinct morphs types for lengths of 4 to 12 phones compared to the 'non Cc' curves (drawn with '+'). The frequency weighted curves (tokens) are very similar with the exception of the 2-phones morphs. With the 'Cc' option, these units are less frequent than without it. Since these small units are more error-prone than longer units, reducing their frequency with the phonetic constraint 'Cc' is promising.

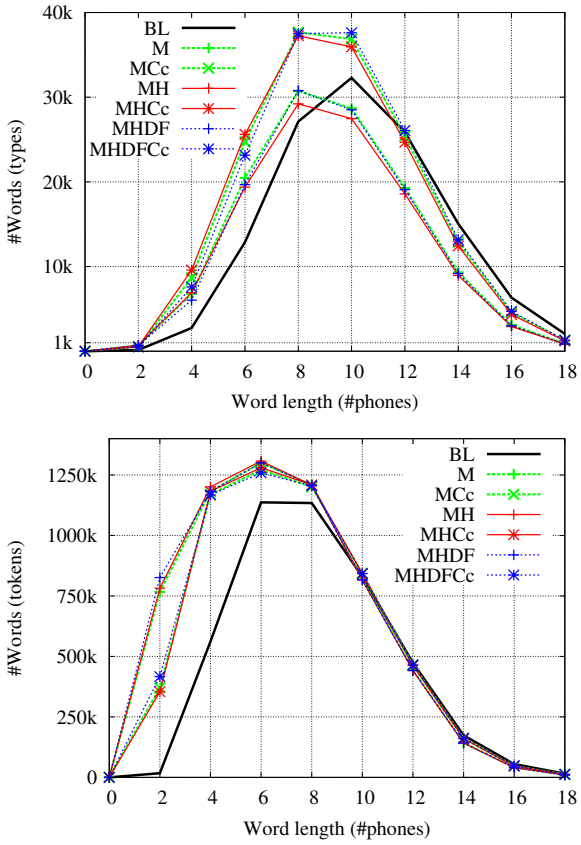

Figure 1: *Number of word types (top) and word tokens (bottom) in the training data as a function of the number of phones for different decomposition options.*

4.3. OOV rates

Table 5 gives the log-likelihood and the type and token OOV rates measured on the development corpus (14.2k words). The language models are Kneser-Ney smoothed four-gram models, and result from the interpolation of two component LMs: one estimated on the web/newspaper texts and the other on the manual transcripts of the audio data. The interpolation coefficient was optimized for each LM by measuring the perplexity on the dev transcripts. Different LMs were built for each set of decompounding options. Since some of the words which are not in the baseline vocabulary are decomposed, the OOV rates are reduced. The relative reduction in OOV rate ranges from 30% to 40% depending on the options. The log-likelihoods are all seen to be smaller than the word-based ones. However, comparing the option sets with and without the phonetic constraint, it can be noted that the use of the confusion constraint increases the likelihood.

5. ASR experiments

This section reports recognition results obtained with systems trained for each of the decomposition option configurations. The baseline system is the word-based system. The speech recognizers all have two decoding passes, with unsupervised acoustic model adaptation after the first pass [13]. Specific acoustic models were built for each option set. All acoustic model sets cover about 10.5k distinct intra-word contexts (3

Table 5: *Likelihood and OOV rates on the dev text. The dev text is comprised of 14.2k tokens and 5.8k types.*

Options	*llh*	*OOV Types*	*OOV Tokens*
BL	-38768	15.7	6.9
M	-41651	10.6	4.3
M Cc	-41126	10.6	4.6
M H	-41733	10.6	4.2
M H Cc	-41408	10.5	4.5
M H DF	-41675	10.6	4.2
M H DF Cc	-41384	10.6	4.5

states per model), with a total of about 8.5k tied states (32 Gaussians/state).

Table 6 gives the word error rates for the different ASR systems, estimated after recombining prefixes and roots back into full words. The full-word baseline system has a WER of 24.0%. The three systems M, M H and M H DF, which do not use the confusion constraint Cc, perform slightly worse than the baseline system. On the contrary, the three Cc systems all give small gains. The worst performance is obtained by the Harris Morfessor algorithm M H, which is the algorithm that split the largest number of words. Nevertheless, the Harris modification seems useful since it produces smaller lexicons than with Morfessor 1.0 and a gain is obtained by adding the Cc option (0.2% absolute with M H Cc compared to M). Concerning the DF option, there is a 0.7% absolute reduction between the M H DF system and its corresponding Cc version. The best performance is obtained with the DF motivated system (M H DF Cc) which achieves a 0.4% absolute improvement compared to the baseline. The confusion constraints between lexical units appears to be useful for identifying recognition units when used in conjunction with word decompounding.

All but 7% of the 971 OOV tokens are decomposed into morphs that are in the recognition dictionary. After recognition 14% of the 902 potentially recognizable decomposed words are recovered (after recomposition). However the difference in WERs of the M and the M H DF Cc systems is about half this number, primarily due to confusion errors of small morphs introduced by the decomposition.

Table 6: *Word Error Rates for the different ASR systems.*

Algorithm Options	*# Morphs*	*WER (%)*
BL	133384	24.0
M	95937	24.1
M Cc	128239	23.9
M H	90740	24.5
M H Cc	126105	23.7
M H DF	94198	24.3
M H DF Cc	128404	23.6

6. Discussion

In this paper, we have described an unsupervised data-driven word decompounding algorithm, which extends the Morfessor algorithm to to better suit speech recognition. The proposed modifications have been validated in recognition experiments, where OOV and WER reductions have been obtained on a less-represented language in which grammatical morphemes are glued to roots.

This algorithm splits words into smaller units in an iterative manner by maximizing a MAP estimate of a lexicon given a word list with frequency counts. The end-of-word probability computation has been modified to allow more splits. A new phonetic-based parameter, motivated by distinctive features and phonetic confusion constraints based on previous audio alignments has been incorporated. We compared systems built with different option sets and the best system incorporates all options. For this system, in comparison to the word-based system, the lexicon size is slightly reduced, and an absolute gain of 0.4% in WER has been achieved. Without the confusion constraints, all systems obtained slightly worse performances than the baseline. For these systems, small units (2 phones long) are very frequent and very error-prone.

The DF parameter is a phonetically motivated parameter. It has been introduced only for vowels and further investigation should be carried out with consonants. In the current implementation, the different terms in the MAP estimate are summed, however it may be useful to weight these terms in order to optimize each contribution. We plan to test the algorithm on another language similar to Amharic, Arabic for instance, as well as on a language in which the word compounding generation process is lexical, such as German or Turkish for example.

7. References

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

[2] Kurimo, M. et al. "Unlimited vocabulary speech recognition for agglutinative languages", Human Language Technology, HLT-NAACL 2006. New York, 2006.

[3] Ordelman, R., van Hessen, A., and de Jong, F. "Compound decomposition in Dutch large vocabulary speech recognition", Eurospeech, Geneva, 2003.

[4] Pellegrini, T., Lamel, L. "Investigating Automatic Decomposition for ASR in Less Represented Languages", Interspeech, Pittsburgh, 2006.

[5] Harris, Z. "From Phoneme to Morpheme", Language **31**:190-222, 1996.

[6] Kiecza, D., Schultz, T. and Waibel, A. "Data-Driven Determination of Appropriate Dictionary Units for Korean LVCSR", ICSP99, pp 323-327, Seoul, August 1999.

[7] Bing Xiang et al. "Morphological Decomposition for Arabic Broadcast News Transcription", ICASSP06, **1**(I):1089-1092, Toulouse, May 2006.

[8] Creutz, M. and Lagus, K. "Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora using Morfessor 1.0", Computer and Information Science, Report A81, Helsinki University of Technology, March 2005.

[9] Halle, M. and Clements, G.N. "Problem Book in Phonology", The MIT Press, 112 pp, 1983.

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

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

[12] Schwartz, R. and al. "Speech recognition in multiple languages and domains: The 2003 BBN/LIMSI EARS system", ICASSP, Montreal, May 2004.

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