Morpheme-Based Language Modeling for Amharic Speech Recognition

Martha Yifiru Tachbelie, Solomon Teferra Abate, Wolfgang Menzel

Department of Informatics, University of Hamburg
Vogt-Kölln-str. 30, D-22527 Hamburg, Germany
tachbeli, abate, menzel@informatik.uni-hamburg.de

Abstract

This paper presents the application of morpheme-based and factored language models in an Amharic speech recognition task. Since using morphemes in both acoustic and language models results, mostly, in performance degradation due to acoustic confusability and since it is problematic to use factored language models in standard word decoders, we applied the models in a lattice rescoring framework. Lattices of 100 best alternatives for each test sentence of the 5k development test set have been generated using a baseline speech recognizer that uses a word-based backoff bigram language model. The lattices have then been rescored with various morpheme-based and factored language models and a slight improvement in word recognition accuracy has been observed.

Keywords: Sub-word based language modeling, Amharic, Lattice rescoring, Factored language modeling

1. Introduction

1.1. Language Modeling

Language models (LM) are fundamental to many natural language applications such as automatic speech recognition (ASR) and statistical machine translation (SMT).

The most widely used kind of language models are statistical ones. They provide an estimate of the probability of a word sequence W for a given task. The probability distribution depends on the available training data and how the context has been defined (Junqua and Haton, 1996). Large amounts of training data are, therefore, required in statistical language modeling so as to ensure statistical significance (Young et al., 2006).

Even if we have a large training corpus, there may be still many possible word sequences which will not be encountered at all, or which appear with a statistically insignificant frequency (data sparseness problem) (Young et al., 2006). There are even individual words that might not be encountered in the training data irrespective of its size (Out of Vocabulary words problem). These problems are more serious for morphologically rich languages.

Morphologically rich languages have a high vocabulary growth rate which results in a high perplexity and a large number of out of vocabulary words (Vergyri et al., 2004). As a solution, sub-word units are used in language modeling [e.g. (Geutner, 1995); (Whittaker and Woodland, 2000); (Byrne et al., 2001); (Kirchhoff et al., 2002) and (Hirsimäki et al., 2005)] to improve the quality of language models and consequently the performance of the applications that use the language models.

1.2. The Morphology of Amharic

Amharic is one of the morphologically rich languages. It is a major language spoken mainly in Ethiopia and belongs to the Semitic branch of the Afro-Asiatic super family. Amharic is related to Hebrew, Arabic and Syrian.

Like other Semitic languages such as Arabic, Amharic exhibits a root-pattern morphological phenomenon. A root is a set of consonants (called radicals) which has a basic 'lexical' meaning. A pattern consists of a set of vowels which are inserted (intercalated) among the consonants of a root to form a stem. The pattern is combined with a particular prefix or suffix to create a single grammatical form (Bender et al., 1976) or another stem (Baye, 2007). For example, the Amharic root sbr means 'break'. By intercalating the pattern ä_ä and attaching the suffix -ä we get säbbärä 'he broke' which is the first form of a verb (3rd person masculine singular in past tense, as in other semitic languages) (Bender et al., 1976). In addition to this non-concatenative morphological feature, Amharic uses different affixes to create inflectional and derivational word forms.

Some adverbs can be derived from adjectives. Nouns are derived from other basic nouns, adjectives, stems, roots, and the infinitive form of a verb by affixation and intercalation. For example, from the noun lIğğ 'child' another noun lIğnät 'childhood'; from the adjective däg 'generous' the noun dägnät 'generosity'; from the stem sInIf, the noun sInIfna 'laziness'; from root qld, the noun qäld 'joke'; from infinitive verb mäsIbär 'to break' the noun mäsIbäriya 'an instrument used for breaking' can be derived. Case, number, definiteness, and gender marker affixes inflect nouns.

Adjectives are derived from nouns, stems or verbal roots by adding a prefix or a suffix. For example, it is possible to derive dIngayama 'stony' from the noun dIngay 'stone'; zIngu 'forgetful' from the stem zIng; sänäf 'lazy' from the root snf by suffixation and intercalation. Adjectives can also be formed through compounding. For instance, hodäsäfi 'tolerant, patient', is derived by compounding the noun hod 'stomach' and the adjective säfi 'wide'. Like nouns, adjectives are inflected for gender, number, and case (Baye, 2007).

Unlike the other word categories such as noun and adjectives, the derivation of verbs from other parts of speech is not common. The conversion of a root to a basic verb stem requires both intercalation and affixation. For instance, from the root gdl 'kill' we obtain the perfective verb stem gäddäl- by intercalating the pattern ä_ä. From this perfective stem, it is possible to derive a passive (tägäddäl-) and a causative stem (asgäddäl-) using the prefixes tä- and as-, respectively. Other verb forms are also derived from roots in a similar fashion.

Verbs are inflected for person, gender, number, aspect, tense and mood (Baye, 2007). Other elements like negative markers also inflect verbs in Amharic.

From the above brief description of Amharic morphology it can be seen that Amharic is a morphologically rich language. It is this feature that makes the development of language models for Amharic challenging. The problems posed by Amharic morphology to language modeling were illustrated by Solomon (2006) who, therefore, recommended the development of sub-word based language models for Amharic.

To this end, Martha and Menzel (2007) and Martha and Menzel (forthcoming) have developed various morpheme-based language models for Amharic and gained a substantial reduction in perplexity and the out-of-vocabulary rate. They have concluded that, in this regard, using sub-word units is preferable for the development of language models for Amharic. However, the Amharic sub-word language models have not been applied to any natural language application, and, therefore, nothing is known whether these language models really bring improvement in the performance of an application or not.

In this study, we applied sub-word (morpheme) based language models to Amharic speech recognition.

Pellegrini and Lamel (2006) have investigated the application of automatic word decompounding (using Harris algorithm) for automatic speech recognition of less-represented languages, specifically Amharic. In their study, the units obtained through decomposition have been used in both acoustic and language models. Word error rate reduction over the base line word-based system has been reported using 2 hours of training data in speech recognition. However, decompounding lexical units with the same algorithm led to worse performance when more training data (35 hours) is used (Pellegrini and Lamel, 2007). This can be explained by a higher acoustic confusability quite similar to other languages [e.g. (Geutner, 1995); (Whittaker and Woodland, 2000) and (Byrne et al., 2001)]. Pellegrini and Lamel (2007) tried to solve this problem by using other modified decompounding algorithms. Their starting algorithm is morfessor (Creutz and Lagus, 2005) which has been modified by adding different information. They were able to achieve a word error rate reduction only when a phonetic confusion constraint was used to hinder decomposition of words which would result in acoustically confusable units.

Unlike Pellegrini and Lamel (2006) and Pellegrini and Lamel (2007), we used morphemes only for the language modeling component to avoid the influence of acoustic confusability on the performance of the speech recognizer. A lattice rescoring framework, as in Whittaker and Woodland (2000) and Kirchhoff et al. (2003), has been applied. Lattices have been generated in a single pass recognition using a bigram word-based language model since the HTK decoder, namely Hvite, does not allow to use higher order n-gram models. The lattices are subsequently rescored using sub-word language models.

In addition, since factored language models (Kirchhoff et al., 2003) enable us to integrate any kind of information that helps to get robust probability estimates of words, we also developed factored language models for Amharic and applied them to a speech recognition task in the same manner the sub-word language models have been applied. Section two describes the baseline speech recognition system and its performance. In section three, we present the morpheme-based and factored language models that we have developed and section four presents the lattice rescoring experiment results. Before that we give a brief introduction to factored language models.

1.3. Factored Language Modeling

Factored language models (FLM) have been first introduced in Kirchhoff et al. (2002) for incorporating various morphological information in Arabic language modeling. In an FLM a word is viewed as a bundle or vector of K parallel factors, that is, $w_n \equiv f^1_n, f^2_n, \dots, f^k_n$. The factors of a given word can be the word itself, stem, root, pattern, morphological classes, or any other linguistic element into which a word can be decomposed. The goal of an FLM is, therefore, to produce a statistical model over these factors. There are two important points in the development of FLM: choosing the appropriate factors which can be done based on linguistic knowledge or using a data driven technique and finding the best statistical model over these factors.

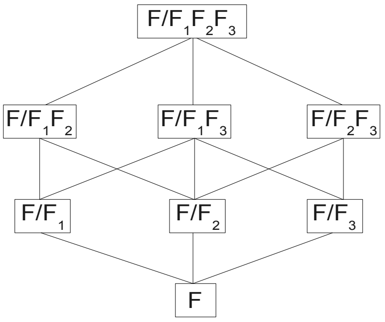

Fig. 1. Possible backoff paths

Unlike normal word or morpheme-based language models, in FLM there is no obvious natural backoff order. In a trigram word based model, for instance, we backoff to a bigram if a particular trigram sequence is not observed in our corpus by dropping the most distant neighbor, and so on. However, in FLM the factors can be temporally equivalent and it is not obvious which factor to drop first during backoff. If we consider a quadrogram FLM and if we drop one factor at a time, we can have six possible backoff paths as it is depicted in Figure 1 and we need to choose a path that results in a better model. Therefore, choosing a backoff path is an important decision one has to make in FLM. There are three possible strategies for deciding on a backoff path: 1) Choosing a fixed path based on linguistic or other reasonable knowledge; 2) Generalized all-child backoff where multiple backoff paths are chosen at run time; and 3) Generalized constrained-child backoff where a subset of backoff paths is chosen at run time (Kirchhoff, Bilmes and Duh, 2008). A genetic algorithm for learning the structure of a factored language model has been developed by Duh and Kirchhoff (2004).

2. The Baseline Speech Recognition System

2.1 Speech and Text Corpus

The speech corpus used to develop the speech recognition system is a read speech corpus developed by Solomon, Menzel and Bairu (2005). It contains 20 hours of training speech collected from 100 speakers who read a total of 10850 sentences (28666 tokens). Compared to other speech corpora that contain hundreds of hours of speech data for training, for example, the British National Corpus (1,500 hours of speech), our models obviously suffer from a lack of training data.

Moreover, the pronunciation dictionary has not been developed by linguistic experts. Encoding of the pronunciation of a corpus can range from very simple and achievable with automatic procedures to very complex and time-consuming that requires manual work with high linguistic expertise. The Amharic speech corpus has been encoded by means of a simple procedure that takes advantage of the orthographic representation which is a consonant vowel syllable.

Although the corpus includes four different test sets (5k and 20k both for development and evaluation), for the purpose of the current investigation we have generated the lattices only for the 5k development test set, which includes 360 sentences read by 20 speakers.

The text corpus used to train the baseline backoff bigram language model consists of 77,844 sentences (868929 tokens or 108523 types).

2.2. The Acoustic and Language Model

The acoustic model is a set of intra-word triphone HMM models with 3 emitting states and 12 Gaussian mixtures that resulted in a total of 33,702 physically saved Gaussian mixtures. The states of these models are tied, using decision-tree based state-clustering that reduced the number of triphone models from 5,092 logical models to 4,099 physical ones.

The baseline language model is a closed vocabulary (for 5k) backoff bigram model developed using the HTK toolkit. The absolute discounting method has been used to reserve some probabilities for unseen bigrams where the discounting factor, D, has been set to 0.5, which is the default value in the HLStats module. The perplexity of this language model on a test set that consists of 727 sentences (8337 tokens) is 91.28.

2.3. Performance of the Baseline System

We generated lattices from the 100 best alternatives for each sentence of the 5k development test set using the HTK tool and decoded the best path transcriptions for each sentence using the lattice processing tool of SRILM (Stolcke, 2002). Word recognition accuracy of this system was 91.67% with a language model scale of 15.0 and a word insertion penalty of 6.0. The better performance (compared to the one reported by Solomon (2006), 90.94%, using the same models and on the same test set) is due to the tuning of the language model and word insertion penalty scales.

3. Morpheme-based and Factored Language Models

3.1. Morpheme-based Language Models

We have developed several sub-word based and factored language models for Amharic using the same data that has been used to develop the baseline language model.

Both statistical and linguistic morphs have been used as units in language modeling. Since there is no morphological analyzer (for Amharic) specifically designed for our purpose, we used a language independent, unsupervised morphology learning algorithm, morfessor (Creutz and Lagus, 2005) to get the statistical morphs. However, this algorithm segments a word only into a sequence of morphemes, and can not extract the root and pattern morphemes of Amharic. On the other hand, a good and complete segmentation of words into morphs leads to a better language model (Martha and Menzel, forthcoming). Therefore, we also investigated the performance of linguistic morpheme-based language models for speech recognition.

The linguistic morphs are obtained according to a manually segmented collection of 72,428 word types (Martha and Menzel, forthcoming). That is, we substituted each word in the corpus with its segmentation if the word is in the manually segmented word collection. Otherwise, the word is left unsegmented. Due to the simplicity of this approach, a substantial share of words (12.3%) in our training data could not be segmented at all.

We developed various sub-word language models using the statistical and linguistic morphs as units in language modeling. We tried to develop ngram language models of order two to four. In all cases we used the SRILM toolkit to train the language models. We smoothed the language models using modified Kneser-Ney smoothing which is known for its state-of-the-art performance unless it became impossible to use it because of zero count of counts. Table 1 presents the perplexity of the various morpheme-based language models.

Language models	**Perplexity**
Linguistic morph bigram	36.55
Linguistic morph trigram	23.09
Linguistic morph quadrogram	18.39
Statistical morph bigram	114.92
Statistical morph trigram	71.61
Statistical morph quadrogram	64.22

Table 1. Perplexity of Morpheme-Based Language Models.

3.2. Amharic Factored Language Models

The manually segmented data has also been used to obtain a factored version of the corpus. Each word is considered a bundle of features including the word itself, part of speech (POS) tag of the word, prefix, root, pattern and suffix. Although words can have more than one prefix or suffix, we considered each word as having zero or one prefix and/or suffix by concatenating a sequence of affixes into a single unit. This corpus has then been used to train various kinds of factored language models.

We developed factored language models (with two and four parents) for which the estimation of the probability of each word depends on the previous word/s and its/their POS, since knowing the POS of a word can tell us which words are likely to occur in its neighborhood (Jurafsky and Martin, 2008).

We also developed a factored language model that considered all the available factors (word, POS, prefix, root, pattern and suffix) as histories and that uses a fixed backoff path by dropping suffix first, then pattern, and so on.

It is difficult to determine which factor combination and which backoff path would result in a robust model yielding an improvement of speech recognition. Therefore, we used the genetic algorithm (Duh and Kirchhoff, 2004) to find the optimal one. The best model is the one that uses four factors (word, prefix, root and pattern) as histories and combines generalized all-child and constrained-child backoff. We applied the two best (in terms of perplexity) models, that differ only in the backoff path, to the speech recognition task. The perplexities of the factored language models are given in Table 2.

Language Models	**Perplexity**
FLM with two parents	115.89
FLM with four parents	17.03
FLM with fixed backoff	97.78
1st Best factor combination	116.41
2nd Best factor combination	192.86

Table 2. Perplexity of Factored Language Models

4. Lattice Rescoring Experiment

The lattices generated as indicated in section two have been rescored using the various morpheme-based language models and decoded to find the best path. An improvement in word recognition accuracy (WRA) has been observed (see Table 3). All morph-based models brought a slight improvement in WRA. However, the linguistic morphs contribute more to the performance improvement than the statistical morphs (an absolute 0.25% increase in accuracy with the linguistic morph trigram model). Using higher order ngram brings only a slight improvement in performance, from 91.77 to 91.82 and then to 91.85 as a result of using trigram and quadrogram language models, respectively.

Since it is problematic to use factored language models in standard word decoders, we substituted each word in the lattice with its factored representation. A word bigram model that is equivalent to the baseline word bigram language model has been trained on the factored data and used as a baseline system for factored representations. This language model has a perplexity of 58.41. The best path transcription decoded using this language model has a WRA of 91.60%, which is slightly lower than the performance of the normal baseline speech recognition system (91.67%). This might be due to the smoothing technique applied in the development of the language models. Although absolute discounting with the same discounting factor has been applied to both bigram models, the unigram models have been discounted differently. While in the baseline word based language model the unigram models have not been discounted at all, in the equivalent factored model the unigrams have been discounted using Good-Turing discounting technique which is the default discounting technique in SRILM.

Language Models Used	**Word Accuracy in %**
Baseline word-based (BL)	91.67
BL + Statistical morph bigram	91.77
BL + Statistical morph trigram	91.82
BL + Statistical morph quadrogram	91.85
BL + Linguistic morph bigram	91.87
BL + Linguistic morph trigram	91.92
BL + Linguistic morph quadrogram	91.89

Table 3. WRA Improvement with Morpheme-based Language Models

The various factored language models (described in section 3.2) have been used to rescore the lattices and brought a considerable improvement in WRA. As it can be seen from Table 4, already one extra information, namely POS, makes language models more robust and consequently the language model improved word recognition accuracy (from 91.60 to 92.92). Although using higher order ngram models brought a slight improvement for the morpheme-based language models, this is not the case for factored language models. The first best factored language model learned by the genetic algorithm outperformed the second one and the factored model that uses all the factors as histories and a fixed backoff path.

Language Models Used	**Word Accuracy in %**
Baseline word bigram (FBL)	91.60
FBL + FLM with two parents	92.92
FBL + FLM with four parents	92.75
FBL + FLM with fixed backoff	92.68
FBL + 1st Best factor combination	92.85
FBL + 2nd Best factor combination	92.50

Table 4. WRA Improvement with Factored Language Models

These results also show that a reduction in perplexity of the language models does not always lead to an improvement in WRA.

5. Conclusion

Several language models (statistical and linguistic morpheme-based and FLMs) have been applied to an Amharic speech recognition task in a lattice rescoring framework. Lattices consisting of the 100 best alternatives for each test sentence have been generated and subsequently rescored with various language models. A considerable improvement in WRA has been observed as a result of using factored language models. The morpheme-based language models brought a slight improvement in WRA. The linguistic morph-based language models contributed more to the performance improvement than the statistical morph-based ones even though a substantial share of the words have been left unsegmented. Therefore, we conclude that morpheme-based language models and factored language models are better suited for Amharic speech recognition than word-based ones.

References

Baye Yimam. (2007). "*yäamarINa säwasäw*", (2nd ed.). Addis Ababa: EMPDE.

Bender, M., Bowen, J., Cooper, R., and Ferguson, C. (1976). *Languages in Ethiopia*. London: Oxford Univ. Press.

Byrne, W., Hajiˇc, J., Ircing, P., Jelinek, F., Khudanpur, S., Krebc, P. and Psutka, J. (2001). On large vocabulary continuous speech recognition of highly inflectional language – Czech. In P. Dalsgaard, B. Lindberg, H. Benner (Eds.) *European Conference on Speech Communication and Technology, Aalborg, Denmark, September 2001*, pp. 487-489.

Creutz, M. and Lagus, K. (2005). *Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.1.* Neural Networks Research Center, Helsinki University of Technology, Tech. Rep. A81.

Duh, K. and Kirchhoff, K. (2004). Automatic learning of language model structure. *In Proceeding of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, August 23-27, 2004,* pp. 148-154.

Geutner, P. (1995). Using morphology towards better large-vocabulary speech recognition systems. *In Proceedings of IEEE 1995 International Conference on Acoustics, Speech and Signal Processing, ICASSP 95, Detroit, Michigan, May 9-12, 1995,* pp. 445-448.

Hirsimäki, T., Creutz, M., Siivola, V. and Kurimo, M. (2005). Morphologically motivated language models in speech recognition. In T. Honkela, V. Könönen, M. Pöllä, O. Simula (Eds.) *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, AKRR'05, Espoo, Finland, June 2005,* pp. 121-126.

Junqua, J.-C. and Haton, J.-P. (1996). *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. London: Kluwer Academic.

Jurafsky, D. S. and Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd ed. New Jersey: Prentice Hall.

Kirchhoff, K., Bilmes, J. and Duh, K. (2008). *Factored language models - a tutorial*. Dept. of Electrical Eng., Univ. of Washington, Tech. Rep.

Kirchhoff, K., Bilmes, J., Henderson, J., Schwartz, R., Noamany, M., Schone, P., Ji, G., Das, S., Egan, M., He, F., Vergyri, D., Liu, D. and Duta, N. (2002). *Novel speech recognition models for Arabic*. Johns-Hopkins University Summer Research Workshop, Tech. Rep.

Kirchhoff, K., Bilmes, J., Das, S., N. Duta, M. Egan, G. Ji, F. He, Henderson, J., Liu, D., Noamany, M., Schone, P., Schwartz, R. and Vergyri, D. (2003). Novel approaches to Arabic speech recognition: Report from the 2002 Johns-Hopkins summer workshop. *In Proceedings of IEEE 2003 International Conference on Acoustics, Speech and Signal Processing, ICASSP 03, Hong Kong, China, April 6-10, 2003,* pp. 344-347.

Martha Yifiru Tachbelie and Menzel, W. (2007). Sub-word based language modeling for Amharic. In G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, N. Nikolov (Eds). *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP-2007, Borovets, Bulgaria, September 27-29, 2007,* pp. 564-571.

Martha Yifiru Tachbelie and Menzel, W. (forthcoming). Morpheme-based Language Modeling for Inflectional Language – Amharic. In N. Nicolov, G. Angelova, R. Mitkov (Eds.) *Recent Advances in Natural Language Processing V (*Book series: *Current Issues in Linguistic Theory)*, Amsterdam and Philadelphia: John Benjamin's Publishing, pp. 301-310.

Pellegrini, T. and Lamel, L. (2006). Investigating automatic decomposition for ASR in less represented languages. *In INTERSPEECH-2006, Pittsburgh, PA, USA, September 17-21, 2006,* pp. 1776-1779.

Pellegrini, T. and Lamel, L. (2007). Using phonetic features in unsupervised word decompounding for ASR with application to a less-represented language. *In INTERSPEECH-2007, Antwerp, Belgium, August 27-31,2007,* pp. 1797-1800.

Solomon Teferra Abate, Menzel, W. and Bairu Tafila. (2005). An Amharic speech corpus for large vocabulary continuous speech recognition. *In INTERSPEECH-2005, Lisbon, Portugal, September 4-8, 2005,* pp. 1601-1604.

Solomon Teferra Abate. (2006). *Automatic Speech Recognition for Amharic*. Ph.D. Dissertation, Univ. of Hamburg.

Stolcke, A. (2002). SRILM — an extensible language modeling toolkit. *In ICSLP-2002, Denber, Colorado, USA, September 16-20, 2002,* pp. 901-904.

Vergyri, D., Kirchhoff, K., Duh, K and Stolcke, A. (2004). Morphology-based language modeling for arabic speech recognition. *In ICSLP-2004, Jedu Island, Korea, October 4-8, 2004,* pp. 2245-2248.

Whittaker, E. and Woodland, P. (2000). Particle-based language modeling. *In ICSLP-2000, Beijing, China, October 16-20, 2000,* pp. 170-173.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P. (2006). *The HTK Book*. Cambridge University Engineering Department.