Capturing Word-level Dependencies in Morpheme-based Language Modeling

Martha Yifiru Tachbelie, Wolfgang Menzel

University of Hamburg
Natural Language System Group
tachbeli,menzel@informatik.uni-hamburg.de

Abstract

Morphologically rich languages suffer from data sparsity and out-of-vocabulary words problems. As a result, researchers use morphemes (sub-words) as units in language modeling instead of full-word forms. The use of morphemes in language modeling, however, might lead to a loss of word level dependency since a word can be segmented into 3 or more morphemes and the scope of the morpheme n-gram might be limited to a single word. In this paper we propose the use of roots to capture word-level dependencies in Amharic language modeling. Our experiment shows that root-based language models are better than the word based and other factored language models when compared on the basis of the probability they assign for the test set. However, no benefit has been obtained (in terms of word recognition accuracy) as a result of using root-based language models in a speech recognition task.

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 language models are statistical language models. 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). Even individual words might not be encountered in the training data irrespective of its size (Out-of-Vocabulary words problem).

The data sparseness problem in statistical language modeling is more serious for languages with a rich morphology. These languages have a high vocabulary growth rate which results in 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 to improve the quality of language models and consequently the performance of the applications that use the language models (Geutner, 1995; Whittaker and Woodland, 2000; Byrne et al., 2001; Kirchhoff et al., 2003; Hirsimäki et al., 2005).

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 (Yimam, 2000EC). For example, the Amharic root sbr means 'break', when we intercalate the pattern ä_ä and attach 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 qälId '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 dInIgayama 'stony' from the noun dInIgay 'stone'; zInIgu 'forgetful' from the stem zInIg; 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 (Yimam, 2000EC).

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 pattern ä_ä. From this perfective stem, it is possible to derive a passive (tägäddäl-) and a causative stem (asgäddäl-) using 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 (Yimam, 2000EC). 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.

1.3. Language modeling for Amharic

Since Amharic is a morphologically rich language, it suffers from data sparseness and out of vocabulary words problems. The negative effect of Amharic morphology on language modeling has been reported by Abate (2006), who, therefore, recommended the development of sub-word based language models for Amharic.
To this end, Tachbelie and Menzel (2007) and Tachbelie and Menzel (2009) have developed various morpheme-based language models for Amharic and gained a substantial reduction in 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. In their experiment, Tachbelie and Menzel (2007) and Tachbelie and Menzel (2009) considered individual morphemes as units of a language model. This, however, might result in a loss of word level dependencies since a word might be segmented into 3 or more morphemes and the span of the n-grams might be limited to a single word. The easiest way of handling this problem might be using higher order n-grams which, however, highly increases the complexity of language models. Therefore, approaches that capture word level dependencies are required to model the Amharic language. Kirchhoff et al. (2003) introduced factored language models that can capture word level dependency while using morphemes as units in language modeling.
In this paper, we present how we captured word-level dependency in morpheme-based language modeling (in the framework of factored language modeling) by taking advantage of the nature of the Amharic language that root consonants represent the lexical meaning of the words derived from them. Section 2. gives a description of our approach for handling word-level dependencies in morpheme-based language modeling and Section 3. presents detail of the language modeling experiment and the results. The language models have been applied to a speech recognition task. Section 4. deals with the speech recognition experiments we have conducted. Conclusions and future research directions are given in Section 5. But before that we give a brief description of factored language modeling.

1.4. Factored language modeling

Factored language models (FLM) have first been introduced in Kirchhoff et al. (2002) for incorporating various morphological information in Arabic language modeling. In FLM a word is viewed as a bundle or vector of K parallel factors, that is, $w_n \equiv f_n^1, f_n^2, ..., f_n^k$. 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 idea is that some of the feature bundles (for example: roots, patterns and morphological class) can uniquely define the words i.e. $(W = w_i) \equiv (R = r_i, P = p_i, M = m_i)$. Therefore, the word n-gram probabilities can be defined in terms of these features/factors as follows (Kirchhoff et al., 2003):

$$\begin{aligned} &P(w_i|w_{i-1}, w_{i-2}) \\ &\quad = P(r_i, p_i, m_i|r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \\ &\quad = P(r_i|p_i, m_i, r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \\ &\qquad P(p_i|m_i, r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \\ &\qquad\quad P(m_i|r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \end{aligned} \quad (1)$$

2. Capturing word-level dependency

As it has been indicated above, the idea in factored language modeling is that some of the features might uniquely define the words and language models can be estimated on the basis of these features. Since in Amharic the root consonants from which a word is derived represent the basic lexical meaning of the word, we considered the root consonants as features that uniquely define the word i.e.

$$w_i \equiv r_i \quad (2)$$

Where w_i is the ith word and r_i a root from which the word is derived. The word n-gram probabilities can, therefore, be defined (according to equation 1) in terms of the consonantal roots as follows.

$$P(w_i|w_{i-1}w_{i-2}) = P(r_i|r_{i-1}r_{i-2}) \quad (3)$$

As it is clear from formula 3, we actually developed a root-based n-gram model to capture word-level dependencies. One can also consider the model as a skipping one since we skip other morphemes during language model estimation.

3. Language modeling experiment

3.1. Morphological analysis

To use morphemes in language modeling a word parser, which splits word forms into their morpheme constituents, is needed. Different attempts (Bayou, 2000; Bayu, 2002; Amsalu and Gibbon, 2005) have been made to develop a morphological analyzer for Amharic using different methods. However, none of the systems can be directly used for our purpose. The systems developed by Bayou (2000) and Bayu (2002) suffer from lack of data. The morphological analyzer developed by Amsalu and Gibbon (2005) seems to suffer from a too small lexicon. It has been tested on 207 words and analyzed less than 50% (75 words) of them.
An alternative approach might be unsupervised corpus-based methods (such as Morfessor (Creutz and Lagus, 2005)) that do not need annotated data. These methods are particularly interesting for resource scarce languages like Amharic. However, the corpus-based morphology learning tools try to find only the concatenative morphemes in a given word and are not applicable for the current study in which the roots (the non-concatenative morphemes) are required. Therefore, we manually segmented 72,428 word types found in a corpus of 21,338 sentences.
To do the segmentation, we used two books as manuals: Yimam (2000EC) and Bender and Fulass (1978). Baye's book describes how morphemes can be combined to form

words. The list of roots in Bender and Fulass (1978) helped us to cross check the roots that we suggest during the segmentation.

However, we do not claim that the segmentation is comprehensive. Since a word type list has been used, there is only one entry for polysemous or homonymous words. For example, the word t'Iru might be an adjective which means 'good' or it might be an imperative verb which has a meaning 'call' (for second person plural). Consequently, the word has two different segmentations. Nevertheless, we provided only one segmentation based on the most frequent meaning of the word in our text. In other words we disambiguated based on the frequency of use in the text. Because the transcription system does not indicate geminated consonants, the geminated and non-geminated word forms, which might have distinct meanings and segmentations, have also been treated in the same manner as the polysemous or homonymous ones. For instance, the word ?at'ägäbu can be treated as an adverb which means 'next to him' or as a verb with a meaning 'they made somebody else full or they satisfied somebody else' based on the gemination of the consonant g. Consequently, this word could have, therefore, been segmented in two different ways: [?at'ägäb + u] if it is an adverb or [?a + t'gb + aa + u] if it is a verb derived from the root t'gb.

3.2. Factored data preparation

As our aim is to handle word level dependencies in the framework of factored language modeling we need a corpus in which each word is represented as a vector of factors or features. Although only the root consonants are required for the present work, we prepared the data in such a way that it can also be used for other experiments. In our experiment each word is considered as a bundle of features including the word itself, part-of-speech tag of the word, prefix, root, pattern and suffix. Each feature in the feature vector is separated by a colon (:) and consists of a tag-value pair. In our case the tags are: W for word, POS for Part-of-Speech, PR for prefix, R for root, PA for pattern and SU for suffix. A given tag-value pair may be missing from the feature bundle. In this case, the tag takes a special value 'null'.

The manually segmented data that include 21,338 sentences or 72,428 word types or 419,660 tokens has been used to prepare the factored version of the corpus. The manually segmented word list has been converted to a factored format and the words in the corpus have been automatically substituted with their factored representation. The resulting corpus has then been used to train and test the root-based language models presented below.

3.3. The language models

Although all the verbs are derived from root consonants, there are words in other part-of-speech class which are not derivations of root consonants. Normally, these words have the value 'null' for the root feature (the R tag). If we consider the word as being equivalent with its root, these words will be excluded from our model which in turn will have a negative impact on the quality of the language models. Preliminary investigation also revealed the fact that the perplexities of the models has been influenced by the null values for the root tag in the data. Therefore, we modified equation 2 as follows.

$$w_i \equiv \begin{cases} r_i, & if\ root \neq null \\ stem_i, & otherwise \end{cases} \quad (4)$$

Where $stem_i$ is the stem of a word (that is not derived from root consonants) after removing all the prefixes and suffixes. We did not introduce a new feature called stem to our factored data representation. But when the word is not a derivation of consonantal root, we consider the stem as a root instead of assigning a null value.

We divided the corpus into training, development and evaluation test sets in the proportion of 80:10:10. We have trained root-based models of order 2 to 5. Since previous experiments (Tachbelie and Menzel, 2007) revealed the fact that Kneser-Ney smoothing outperforms all other smoothing methods, we smoothed the root-based language models with this technique. Table 1 shows the perplexity of these models on the development test set. A higher improvement in perplexity (278.57 to 223.26) has been observed when we move from bigram to trigram. However, as n increases above 3, the level of improvement declined. This might be due to the small set of training data used. As n increases more and more n-grams might not appear in the training corpus and therefore the probabilities are computed on the basis of the lower order n-grams. The pentagram model is the best model compared to the others. This model has a perplexity of 204.95 on the evaluation test set.

Root ngram	Perplexity	Word ngram	Perplexity
Bigram	278.57	Bigram	1148.76
Trigram	223.26	Trigram	989.95
Quadrogram	213.14	Quadrogram	975.41
Pentagram	211.93	Pentagram	972.58

Table 1: Perplexity of root- and word-based models on development test set

In order to measure the benefit gained from the root-based model, we have developed word based models with the same training data. These models have also been tested on the same test data (consisting of 2,134 sentences or 20,989 words) and the same smoothing technique has also been applied. The difference from the root-based models is that complete words are used as units instead of roots. The pentagram model has a perplexity of 972.58 (as shown in table 1) on the development test set. Moreover, the number of out-of-vocabulary words is much lower (295) in the root-based models than in the word based ones (2,672). Although the test set used to test the word- and root-based models has the same number of tokens, direct comparison of the perplexities of these models is still impossible since they have a different number of out-of-vocabulary words. Therefore, comparison of the best root- and word based models has been performed on the basis of the probability they assign to the test set. The log probability of the best root based model is higher (-53102.3) than that of the word based model (-61106.0). That means the root-based

models are better than the word based ones with respect to the probability they assign to the test set.
Word based language models that use one additional word-dependent feature in the ngram history have also been developed since integrating features, such as part-of-speech, into language models might improve their quality. In these models, a word trigram probability is estimated, for example, as $w_n|w_{n-2}pos_{n-2}w_{n-1}pos_{n-1}$ instead of $w_n|w_{n-2}w_{n-1}$. Table 2 gives the perplexities of these models. Although they are better than the word only models (in terms of the probability they assign to the test set), none of the models outperformed the root-based ones. However, as the levels of detail modeled in root-based and other language models are different, the root-based models have been applied in a speech recognition system to prove that they are really better than the other models.

Language models	Perplexity
W/W2,POS2,W1,POS1	885.81
W/W2,PR2,W1,PR1	857.61
W/W2,R2,W1,R1	896.59
W/W2,PA2,W1,PA1	958.31
W/W2,SU2,W1,SU1	898.89

Table 2: Perplexity of models with different factors

4. Speech recognition experiment

In order to analyse the contribution (in terms of performance improvement) of the root-based and other factored language models in a speech recognition task, the speech recognition system which has been developed by Abate (2006) has been used. Section 4.1. presents the speech recognition system. To make our results comparable, we have developed root-based and other factored language models that are equivalent with the ones elucidated in section 3.3. They differ from the models described in the preceding section by having been trained on the text which has been used to develop the bigram word based language model that was originally used in the speech recognition system. These language models, interpolated with the bigram word based language model, have then been used to rescore lattices generated with the speech recognition system. We applied a lattice rescoring framework, because it is problematic to use factored language models in standard word decoders. Section 4.2. presents the new set of language models and 4.3. deals with the result of the lattice rescoring experiment.

4.1. The speech recognition system

4.1.1. The speech and text corpus

The speech corpus used to develop the speech recognition system is a read speech corpus (Abate et al., 2005). It contains 20 hours of training speech collected from 100 speakers who read a total of 10,850 sentences (28,666 tokens). Compared to other speech corpora that contain hundreds of hours of speech data for training, this corpus is obviously small in size and accordingly the models will suffer from a lack of training data.
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 backoff bigram language model consists of 77,844 sentences (868,929 tokens or 108,523 types).

4.1.2. The acoustic, lexical and language models

The acoustic model is a set of intra-word triphone HMMs 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.
Encoding the pronunciation dictionary 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 pronunciation dictionary has been encoded by means of a simple procedure that takes advantage of the orthographic representation (a consonant vowel syllable) which is fairly close to the pronunciation in many cases. There are, however, notable differences especially in the area of gemination and insertion of the epenthetic vowel.
The 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 (8,337 tokens) is 91.28.

4.1.3. Performance of the 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 (WRA) 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 Abate (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 factors.

4.2. Root-based and factored models

The manually segmented data has also been used to obtain a factored version of the corpus that was used to develop the backoff bigram word based language model. The factored version of the corpus has been prepared in a way similar to the one described in Section 3.2. This corpus has then been used to train closed vocabulary root-based and factored language models. All the factored language models have been tested on the factored version of the test set used to test the bigram word based language model.
We have developed root-based n-gram language models of order 2 to 5. The perplexity of these models on the development test set is presented in Table 3. The highest perplexity

improvement has been obtained when the n-gram order has been changed from bigram to trigram.

Language models	Perplexity	Logprob
Root bigram	113.57	-18628.9
Root trigram	24.63	-12611.8
Root quadrogram	11.20	-9510.29
Root pentagram	8.72	-8525.42

Table 3: Perplexity of root-based models

Other factored language models that take one word feature (besides the words) in the n-gram history have been developed. The additional features used are part-of-speech (POS), prefix (PR), root (R), pattern (PA) and suffix (SU). The models are equivalent in structure with the factored language models described in Section 3.3. The perplexity and log-probability of these models are presented in Table 4. The models are almost similar in perplexity and probability.

Language models	Perplexity	Logprob
W/W2,POS2,W1,POS1	10.614	-9298.57
W/W2,PR2,W1,PR1	10.67	-9322.02
W/W2,R2,W1,R1	10.36	-9204.7
W/W2,PA2,W1,PA1	10.89	-9401.08
W/W2,SU2,W1,SU1	10.70	-9330.96

Table 4: Perplexity of other factored language models

4.3. Lattice rescoring

Since it is problematic to use factored language models in standard word decoders, we substituted each word in the lattices with its factored representation. A word bigram model that is equivalent to the one originally used in the speech recognition system has been trained on the factored data and used for factored representations. This language model has a perplexity of 63.59. 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 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 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.

The root-based and the other factored language models (described in Section 4.2.) have been used to rescore the lattices.

All the factored language models that integrate an additional word feature in the n-gram history brought an improvement in WRA. Models with four parents did not bring much improvement when the maximal n-gram order to be used for transition weight assignment was set to 2. However, when trigrams are used, all the models brought notable improvement (see Table 5).

Language models	Word recognition accuracy in %
Factored word bigram (FBL)	91.60
FBL + W/W2,POS2,W1,POS1	93.60
FBL + W/W2,PR2,W1,PR1	93.82
FBL + W/W2,R2,W1,R1	93.65
FBL + W/W2,PA2,W1,PA1	93.68
FBL + W/W2,SU2,W1,SU1	93.53

Table 5: WRA with other factored language models

Unlike the other factored language models, root based language models led to a reduced word recognition accuracy as Table 6 shows. Although the higher order root-based model, namely the penta-gram, assigned the highest probability to the test set compared to all the other factored language models, it resulted in a WRA which is below that of the original speech recognition system.

Language models	Word recognition accuracy in %
Factored word bigram (FBL)	91.60
FBL + Root bigram	90.77
FBL + Root trigram	90.87
FBL + Root quadrogram	90.99
FBL + Root pentagram	91.14

Table 6: WRA with root-based models

5. Conclusion and future work

In Amharic, the consonantal roots from which a word is derived represent the basic lexical meaning of the words. Taking advantage of this feature, root-based language models have been developed as a solution to the problem of loss of word-level dependencies in Amharic morpheme-based language modeling. Our experiment shows that root-based language models are better than the word based and other factored language models when they are compared on the basis of the probability the models assign to a test set.

Since the best way of comparing language models is applying them to the target application for which they are developed and see whether they bring improvement in the performance of the application or not, the root-based models have been applied to a speech recognition task in a lattice rescoring framework. However, the speech recognition system did not benefit from these models. Thus, other ways of integrating the root-based models to a speech recognition system might be worth exploring.

6. Acknowledgment

We would like to thank the reviewers for their constructive and helpful comments.

7. References

Solomon Teferra Abate, Wolfgang Menzel, and Bairu Tafila. 2005. An Amharic speech corpus for large vocabulary continuous speech recognition. In *Proceedings of 9th. European Conffference on Speech Communication and Technology, Interspeech-2005*.

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

Saba Amsalu and Dafydd Gibbon. 2005. Finite state morphology of amharic. In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, pages 47–51.

Abiyot Bayou. 2000. Developing automatic word parser for amharic verbs and their derivation. Master's thesis, Addis Ababa University.

Tesfaye Bayu. 2002. Automatic morphological analyzer for amharic: An experiment employing unsupervised learning and autosegmental analysis approaches. Master's thesis, Addis Ababa University.

M. L. Bender and H. Fulass. 1978. *Amharic Verb Morphology: A Generative Approach*. Michigan State University, Michigan.

M.L. Bender, J.D. Bowen, R.L. Cooper, and C.A. Ferguson. 1976. *Languages in Ethiopia*. Oxford Univ. Press, London.

W. Byrne, J. Hajič, P. Ircing, F. Jelinek, S. Khudanpur, P. Krebc, and J. Psutka. 2001. On large vocabulary continuous speech recognition of highly inflectional language - czech. In *Proceeding of the European Conference on Speech Communication and Technology*, pages 487–489.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.1. Technical Report A81, Neural Networks Research Center, Helsinki University of Technology.

P. Geutner. 1995. Using morphology towards better large-vocabulary speech recognition systems. In *Proceedings of IEEE International on Acoustics, Speech and Signal Processing*, volume I, pages 445–448.

Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, and Mikko Kurimo. 2005. Morphologically motivated language models in speech recognition. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pages 121–126.

Jean-Claude Junqua and Jean-Paul Haton. 1996. *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Kluwer Academic, London.

Katrin Kirchhoff, Jeff Bilmes, John Henderson, Richard Schwartz, Mohamed Noamany, Pat Schone, Gang Ji, Sourin Das, Melissa Egan, Feng He, Dimitra Vergyri, Daben Liu, and Nicolae Duta. 2002. Novel speech recognition models for arabic. Technical report, Johns-Hopkins University Summer Research Workshop.

Katrin Kirchhoff, Jeff Bilmes, Sourin Das, Nicolae Duta, Melissa Egan, Gang Ji, Feng He, John Henderson, Daben Liu, Mohamed Noamany, Pat Schone, Richard Schwartz, and Dimitra Vergyri. 2003. Novel approaches to Arabic speech recognition: Report from the 2002 johns-hopkins summer workshop. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 344–347.

Andreas Stolcke. 2002. SRILM — an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, volume II, pages 901–904.

Martha Yifiru Tachbelie and Wolfgang Menzel. 2007. Subword based language modeling for Amharic. In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, pages 564–571, September.

Martha Yifiru Tachbelie and Wolfgang Menzel, 2009. *Morpheme-based Language Modeling for Inflectional Language – Amharic*. John Benjamin's Publishing, Amsterdam and Philadelphia.

Dimitra Vergyri, Katrin Kirchhoff, Kevin Duh, and Andreas Stolcke. 2004. Morphology-based language modeling for Arabic speech recognition. In *Proceedings of International Conference on Spoken Language Processing*, pages 2245–2248.

E. Whittaker and P. Woodland. 2000. Particle-based language modeling. In *Proceeding of International Conference on Spoken Language Processing*, pages 170–173.

Baye Yimam. 2000EC. *yäamarIña säwasäw*. EMPDE, Addis Ababa, 2nd. ed. edition.

Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland, 2006. *The HTK Book*. Cambridge University Engineering Department.