International Journal of Hybrid Information Technology
Vol.8, No.8 (2015), pp.327-342
http://dx.doi.org/10.14257/ijhit.2015.8.8.33

Morpheme Segmentation and Concatenation Approaches for Uyghur LVCSR

Mijit Ablimit[1], Tatsuya Kawahara[2], Askar Hamdulla[3]

[1]*Postdoctoral Research Station of Computer Science and Technology, Xinjiang University, Urumqi, China 830046*
[2]*School of Informatics, Kyoto University, Kyoto, Japan*
[3]*School of Software, Xinjiang University, Urumqi, China 830046*
mijit601@gmail.com, kawahara@ar.media.kyoto-u.ac.jp
Corresponding Author: askarhamdulla@gmail.com

Abstract

In this paper, various kinds of sub-word lexica are thoroughly investigated under the framework of Uyghur LVCSR system. Experimental results show that it is inefficient to directly model based on word units or small units like morpheme or even syllable units. It is observed that an optimal sub-word unit set between word and morpheme units can better fit for ASR system. In order to select best unit set we have investigated several effective unit segmentation, concatenation approaches, and their ASR performances. For segmentation approach, we investigate a supervised segmentation which split words into the smallest functional units - the linguistic *morphemes, and an unsupervised segmentation which extract* pseudo-morphemes *(or* statistical morphemes*). In supervised model, a leaning algorithm is trained on a manually prepared training corpus, and morpho-phonetics changes are analyzed. In the unsupervised model, the* Morfessor *tool is used to extract* pseudo-morphemes *from a raw text corpus. For concatenation approach, several effective concatenation approaches are investigated based on linguistic morphemes. First is the data-driven approach which concatenates morpheme sequences based on certain measures like co-occurrence frequency or mutual probability. Second is a* model *based approach which merges units with global statistical criteria. In this study, the* Morfessor *program is revised and turned into concatenation program by controlling segmentation points. Third is the two-layer-lexica based concatenation approach which extracts an optimal sub-word unit set by aligning and comparing the ASR results of word and morpheme two lexical layers. This method utilizes both speech and text, and produced the best results in terms of WER and lexicon size, and proved to be very stable. The best optimal lexicon, which is obtained totally on the basis of HMM based acoustic model, outperformed all other baseline lexica. And when all these lexica are directly incorporated with a deep neural network (DNN) based acoustic model, without changing the speech and text training corpora and language models, the optimal lexicon not only drastically improved the ASR accuracy but also outperformed other units as a proof of the generality of the two-layer-lexica based approach.*

Keywords: *Speech recognition; Uyghur; morpheme; lexicon optimization*

1. Introduction

Uyghur language belongs to Turkish Language Family of Altaic Language system. Words are naturally separated in text. It is an agglutinative language in which words are formed by productive affixation of derivational and inflectional suffixes to a root without any splitter between them. The derivational suffixes make semantic changes, while the inflectional suffixes make syntactic changes. The stem

ISSN: 1738-9968 IJHIT

set, in this paper, is consisted of the roots and some stems which are formed by attaching a derivational suffix to a root.

Words in agglutinative languages are relatively long, and the vocabulary size of these languages is growing up proportionally with the corpus size, causing out-of-vocabulary (OOV) and data sparseness problems. It is timely and spatially inefficient to use words as the basic unit set [1]. Therefore, sub-word units like morphemes are conventionally adopted in many inflectional languages, such as Japanese, Korean, Turkish, Finnish, German and Arabic [2-12]. However, short units shrink the context of statistical models, and prone to morpho-phonetic confusions. When sequence of units are merged or split, unit boundaries are phonetically harmonized in the speech which reflects as the morpho-phonetic changes in the text.

Many language processing tasks including parsing, semantic modeling, information retrieval, and machine translation frequently requires a morphological analysis of the language at hand [14-16]. In this research, the morpheme is mainly investigated as the foundation of concatenative approaches for ASR tasks, for it can provide high coverage, low vocabulary size, acceptable ASR performance, and semantic and syntactic relations. Smaller units like syllables better be phonetic particles, and too short to hold its contextual relations. Words and morphemes have their merits and demerits respectively, the reasons can be explained in both statistical and linguistic ways. Therefore, analyzing these reasons and finding an optimal unit set which has both high coverage and better constraints are very important research topics for highly inflectional languages. An optimal lexicon can better generalize for texts of especially limited resources, and increase the reliability of statistical models [13].

In this paper, we investigate supervised and unsupervised segmentation of morphemes and pseudo-morphemes (statistical morphemes), and their ASR performances. We also investigate several effective concatenative methods, such as data-driven approach, statistical model based approach, and two-layer-lexica based approach, and their ASR performances. Based on the morpheme unit, concatenation approaches does not cross the word boundaries, so that the optimized lexical units are the granules between word and morpheme layers.

The present day typical ASR system consisted of acoustic model (AM) and language model (LM). The AM generates phonetic sequences based on the speech, while the LM generates morpho-syntactic unit sequences based on text. Our experiments are based on the same AM which can map all the morpho-phonetic changes of speech into the text of character sequences in the Uyghur language. These morpho-phonetic changes are extracted and analyzed in our general purpose morphological analyzer. The linguistic information including morpheme and word boundaries are preserved by labeling stems and suffixes. Thus all the sub-word units are conveniently re-merged into words and compared by word-error-rate (WER).

In this paper, various linguistic particle sets are investigated and evaluated based on an Uyghur LVCSR system. All the morphological lexica based ASR systems are separately investigated and compared under a hidden Markov model (HMM) framework and a deep neural network (DNN) framework based on the same text and speech training corpora.

The remainder of the paper is organized as follows: Fist we discuss sub-word segmentation methods in Section 2, then, unit concatenation methods in Section 3. Next, we demonstrate experimental evaluations for segmented and concatenated lexica in Section 4, before concluding in Section 5.

2. Morpheme Segmentation Approaches

Uyghur text is written as pronounced, each phoneme is recorded by a character, total 32 characters for 32 phonemes (8 vowels and 24 consonants). The surface realizations of the morphological structure are constrained and modified by a number of morphological and phonetic changes such as insertion, deletion, phonetic harmony, and disharmony (vowel assimilation, vowel weakening) [1].

There are linguistic morphemes and pseudo-morpheme to be extracted and applied for ASR systems. The morpheme, smallest functional unit, is extracted in a supervised manner. A statistical leaning model can be constructed and trained on a manually prepared corpus. And, the pseudo-morphemes are extracted in an unsupervised manner which can split words into morpheme-like units from a raw text corpus by using a probabilistic criterion [13-14]. These pseudo-morphemes are specially designed for certain applications and not strictly meaningful units.

2.1. Supervised Morpheme Segmentation

Linguistic morphemes have their standard forms and surface forms. There are 1~4 different surface forms for a morpheme in Uyghur language. Various surface forms are the result of phonetic harmony when the units are merged to form longer morphological units. And the strong syllable bond in Uyghur language causes some morphological changes like deletion, insertion, and substitution. A general purpose morpheme segmenter tool must consider morpho-phonetic changes of sub-word units. The morpheme structure of Uyghur words is "prefix + stem + suffix1 + suffix2 +…" A root (or stem) is followed by zero to many suffixes as in Example 1. In this research, 108 suffix types are defined strictly according to their semantic and syntactic functions, which have 305 surface forms. A few words have a (only one) prefix preceding a stem; 7 kinds of prefixes are considered.

A general purpose Uyghur morpheme segmenter has been developed by training a learning model on a manually prepared training corpus. As the training data, a text corpus of 10025 sentences and their manual segmentations are prepared as in Table 1. These sentences are collected from general topics, unrelated. Furthermore, considering the limited size of the training corpus, we prepared more than 30K stems independently and used for the segmentation task to produce a probable segmentation result for an unknown word which is not covered by the training corpus. Considering flexibility of language, we keep various segmentation forms of a same word as in Example 2. For example: "work+ing" is segmented to the root while "worker" is kept as a stem. This will expand the size of stem vocabulary, but may be more convenient for analyzing semantic and syntactic context.

Table 1. Manually Segmented Morpheme Corpus

units	tokens	vocabulary
word	139.0k	35.37k
morpheme	261.7k	11.8k
character	936.8k	
sentence	10025	

(**Example 1** morpheme segmentation)

Müshükning kəlginini korgən chashqan hoduqup qachti.

(The mouse seeing the coming cat was startled and escaped.)

Müshük+ning kəlgən+i+ni kor+gən chashqan hoduq+up qach+ti. (morpheme sequence)

(**Example 2** various segmentations of a word)

oqutquchi (teacher{stem})= oqut(teach){root} + quchi(er) {suffix}

yazghuchi(writer{stem}) = yaz(write) {root}+ghuchi(er) {suffix}

hesablinidu = hesab+la+n+idu, hesab+lan+idu;

The learning model is based on an intra-word bi-gram model as in equation (2-1). For a candidate word, all the possible segmentation results are extracted in reference to both stem and suffixes, and their probabilities are calculated to produce the best option.

$$\begin{cases} \mathbf{P(stem,\ firstSuffix)} \\ \mathbf{P'(stem)P(anySuffix\,|\,stem)} \quad \text{for smoothing} \end{cases}$$
$$\text{in which}$$
$$\mathbf{P'(stem)} = \frac{\mathbf{stemFrequency}}{\mathbf{(stemToken+stemVocabulary)}} \qquad (2\text{-}1)$$
$$\mathbf{P(anySuffix\,|\,stem)} \quad \text{probability of a stem linked with a suffix}$$

In this approach, the identification of stem and word-ending boundary is the most important part. At first, a word is split into two parts, a stem and a word-ending (combined suffix or stem-ending in some papers). Then, the word-ending is re-segmented into singular-suffixes. So the segmenter can perform both stemming and segmenting tasks for general purposes.

For an open test set, the word coverage is 86.85%, the morpheme coverage is 98.44%, and the morpheme segmentation accuracy is 97.66%. This morpheme segmenter can output both on the standard forms and on the surface forms without costing segmentation accuracy [11]. The manually prepared morpheme sequences are defined on their strict linguistic functions. Morpho-phonetic variations are also learnt from the training corpus. Our general purpose morpho-phonetic analyzer can segment Uyghur text into phonemes, syllables, morphemes, and words with high accuracy. And can be applied to different research purposes.

2.1. Unsupervised Morpheme Segmentation

Unsupervised morpheme segmentation approaches extract pseudo-morpheme units from an un-annotated raw text corpus. Extraction of linguistic morphemes is not the main target. The practical purpose of the unsupervised segmentation is to provide a lexicon which is smaller and generalizes better than a vocabulary consisting of words as they appear in text. Such a lexicon could be useful in statistical LM of ASR. Unsupervised probabilistic models are designed either to segment word units into sub-word (or pseudo-morpheme) units, or automatically selecting granules from a text according to a probabilistic distribution [13-14]. Surface forms are unchanged for pseudo-morphemes, since they are not strictly functional units.

Morfessor program developed by Creutz [13] is a popular program for the unsupervised induction of a simple morphology from a raw text data. The main idea of this program is to use frequent word units to segment infrequent words. Pseudo-morphemes are extracted from a raw text corpus in an unsupervised way by using a probabilistic criterion of maximum a posteriori (MAP). Totally based on a raw text corpus, the joint probability of the optimized sub-word unit sequence is maximized. The Morfessor has been successfully applied to ASR of several languages such as Finnish and Turkish [17-22], and reported to have improved ASR performance.

The model of language (M) consists of units and their various properties, and the goal is to find an optimal model which maximizes the following probability.

$$argmax\, P(M|corpus) = argmax\, P(corpus|M)P(M)$$

where

$$P(M) = P(properties) = M!\ P(\ properties(t_1) \ldots properties(t_N)) \quad \textbf{(2-2)}$$

The properties can simply be frequency, length, or some linguistic and phonetic attributes [13].

$$P(M) = M!\ P(\ freq(t_1) \ldots freq(t_N)) \cdot \prod_{i=1}^{N} [(1 - P(\#))^{length(t_i)} \cdot P(\#) \cdot \prod_{j=1}^{length(t_i)} P(c_j^{t_i})] \quad \textbf{(2-3)}$$

In this work, it is assumed that words are consisting of lengthy sequences of segments. This model is suitable for languages with agglutinative morphological structure. And no distinction is made between stems and affixes. However, we add stem and suffix labels in order to conveniently recover ASR results from pseudo-morpheme into words which can be used for fair WER comparison.

3. Morpheme Concatenation Approaches

The general idea of concatenation approach is merging the frequently co-occurred and easily confused units while splitting less frequent and easily misrecognized units without causing much phonetic confusions. There are data-driven methods, statistical model based methods, and two-layer-lexica based approaches investigated in this paper. Figure 1 demonstrates the overall concatenation optimization process.

In these concatenation approaches, certain morphemes are merged into longer units to form a new granule. The new built lexicon is used to build a new LM for ASR system. WER and lexicon size are compared to evaluate every lexicon set. Morpho-phonetic confusions in certain morphemes can be avoided when longer units are formed through concatenation.

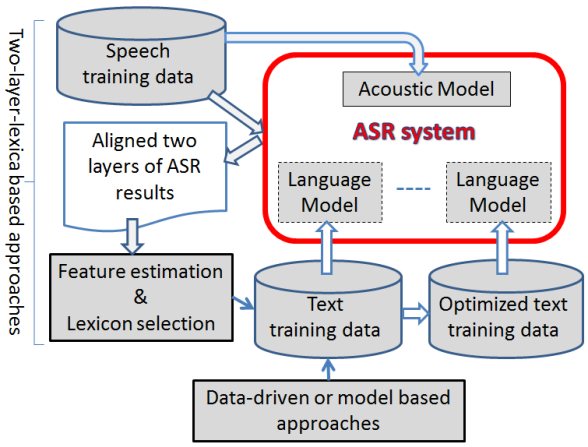

Figure 1. Overall Flow-Chart of Various Lexicon Optimization Approaches

3.1. Data-driven morpheme concatenation approaches

Data-driven approaches merge sequential units based on certain measures like co-occurrence frequency and mutual probability. Below are some widely reported effective data-driven approaches [25].

1) Stem & word-ending. Suffixes in Uyghur language are often very short, one or two phonemes, so can easily be confused. Remerging singular suffixes makes a word consisting of two parts: stem and wording-ending [26]. Thus, word-endings can reduce phonetic alterations between singular suffixes.

2) Co-occurrence frequency. A simple model based on statistical co-occurrence is built by merging frequently co-occurred unit sequences. Specifically, we count unit bigram co-occurrence frequency $C(m_i\ m_j)$, and concatenate them if the frequency is higher than a threshold.

3) Mutual probability. Another statistical measure is mutual probability (MP) [25]. It is calculated as a geometrical mean of forward and reverse bigram probabilities as in equation (3-1). In this method, the pairs of unit counts $C(m_i\ ,m_j)$ must be high while the unigram counts $C(m_i), C(m_j)$ are smaller to prevent the units $m_i\ ,m_j$ to occur in conjunction with other units. This criterion is iteratively applied to morpheme based text corpus.

$$MP(m_i\ m_j) = \sqrt{\mathrm{P_f}(m_i|m_j)\mathrm{P_r}(m_j|m_i)} = \frac{\mathrm{C}(m_i,m_j)}{\sqrt{\mathrm{C}(m_i)\mathrm{C}(m_j)}} \quad (3\text{-}1)$$

3.2. A statistical model based morpheme concatenation approach

A statistical model based concatenation can be developed by maximizing a likelihood function, or by searching a global optimal point for a statistical criterion. An optional model is the Morfessor program which can produce an optimal lexicon from a text of word sequences (section 2.2). If we assume that text is consisted of morphemes instead of characters, the Morfessor program could produce sub-word units which are actually the concatenation of morphemes. Specifically, we can insert morpheme boundaries for word-text corpus by using our supervised morpheme segmenter, and then feed this corpus to Morfessor. When the searching point of Morfessor algorithm is confined to morpheme boundaries only, the output is a kind of regrouping of morpheme sequences within the word boundary. We can see that word and morpheme boundaries are preserved here. In this method, a global optimal point is obtained by maximizing equation (2-2).

3.3. Two-layer-lexica based morpheme concatenation approaches

It is convenient to directly observe the ASR results, and enumerate the problematic patterns, instead of speculating reasons of unknown results. By examining the ASR results of morpheme lexicon, easily confused unit sequences can be systematically extracted and analyzed. Overall view of the proposed scheme is depicted in Figure 1. Baseline ASR systems should be prepared with both morpheme-based lexicon and word-based lexicon, and separately they are applied to decode a large-scale speech data into two layers of unit sequences. A practical method can be aligning and comparing the ASR results of word and morpheme lexica, and extract problematic morpheme sequences or CRITICAL samples as given in Table 2.

By observing the aligned ASR results, CRITICAL samples, “OX” in which the word is correctly recognized while aligned morphemes are misrecognized, can be extracted for further analysis. A simple method which can decrease the WER would be to extract all the problematic morpheme sequences according to their error frequency, merge them into words, and add them to the morpheme lexicon. Our preliminary study showed that this naive method is very effective, but this method only utilized CRITICAL samples (about 28.5%), it is difficult to cover all the erroneous words in the open test data. Therefore, we explore more generalized methods. Furthermore, when these two layers of ASR results

are different, neither of them is correct in most cases (approximately 68%). This majority of samples can be utilized in some sophisticated machine learning methods.

Table 2 Example of ASR results of morpheme and word units

reference (word) (English word-by-word)	Yash cheghinglarda bilim elishinglar kerək young when_you_ are knowledge acquire must
reference (morpheme)	Yash chegh_ing_lar_da bilim el_ish_ing_lar kerək
ASR result (word)	Yash cheghinglarda bilim berishinglar kerək O O O X O
ASR result (morpheme)	Yash chegh_ing_da bilim el_ish_ing_lar kerək O X O O O

CRITICAL

A more sophisticated approach is to extract some morphological features from the aligned samples, for example, length and unigram as the clear cause of misrecognition [26]. These manually selected features can utilize all the aligned samples. A discriminative evaluation function can be designed to discriminatively evaluate the candidate units. And this function can be trained on all the samples extracted from the aligned two-layer ASR results as in Table 2. We feed all the samples to the leaning algorithm which can decide the best concatenation options. In this case, we can only use the two-layer ASR results and no need to compare with the transcription of training speech corpus [23]. This property of not using the reference transcripts makes this approach an unsupervised learning, so that we can make use of enormous un-transcribed speech data.

1) Evaluation functions for lexical features

In this scheme, each word w is described by a set of features ($\Phi_s(w), s = 1, \dots k$) of the constitute morphemes ($w = m_1 m_2 \dots$), and its desired value y defined by the differences of ASR results of the aligned two layers of units. We assume that they are binary (1 for true, 0 or -1 for otherwise).

Given all the training pairs extracted from two-layer ASR results, $(\Phi(w_i), y_i), i = 1, \dots l,\ \Phi(w_i) \epsilon \{-1, 1\},\ y_i \epsilon \{-1, +1\}$, we feed them to the learning scheme which could be Perceptron or SVM. In this study, we investigate the SVM machine learning algorithm which is more robust for outlier samples [27].

For SVM, we adopt a linear binary classifier [27]. When the word is misrepresented by the morpheme sequence, the desired value is $y_i = +1$, otherwise $y_i = -1$. Given a set of training sample pairs $(\boldsymbol{\Phi}(w_i), y_i)$, this method solve the following unconstrained optimization problem with a loss functions $\xi(\alpha;\ \boldsymbol{\Phi}(w_i), y_i)$:

$$min_\alpha \frac{1}{2} \alpha^T \alpha + C \sum_{i=1}^{l} \xi(\alpha;\ \boldsymbol{\Phi}(w_i), y_i) \quad \textbf{(3-2)}$$

$$\xi(\alpha;\ \boldsymbol{\Phi}(w_i), y_i) = \max(1 - y_i \alpha^T \Phi(w_i), 0)^2 \quad \textbf{(3-3)}$$

where C>0 is a penalty parameter. The SVM optimization is stopped at the tolerance of 0.1[27].

Training feature sample pairs $(\Phi(w_i), y_i)$ are extracted independently for every word with its aligned morpheme sequence. This model evaluate every word according to its

features $\Phi(w_i)$, which indicates the potential importance of the word (or sub-word) to be included in the lexicon, or how likely WER will be reduced by adding this new entry. Note that these models can be used for any words or even sub-words consisting of morphemes.

2) Morpheme bigram feature

In this paper, we focus on typical morpheme entries and their bigram patterns. The bigram feature can capture the context, proved to be most effective feature [23]. Each bigram feature of morpheme bigram $(m_i\ m_j)$ is defined from an aligned sample within a word w as in equation (3-4). A specific weight α_s is estimated for each bigram entry.

$$\Phi_{\text{bigram\_}m_i m_j}(w) = \begin{cases} 1 \text{ if } bigram\ (m_i\ m_j)\ exists\ in\ w \\ 0 \text{ otherwise} \end{cases} \quad \textbf{(3-4)}$$

Below is a feature of the morpheme bigram $(\_ing\ \_lar)$ in the aligned sample of word $(cheghinglarda)$ from Table 2.

$$\Phi_{\text{bigram\_}ing\_lar}(cheghinglarda) = 1$$

3) Weight estimation with discriminative learning

The weight α_s is estimated for every bigram feature, based on corresponding desired output y_i. The desired value is defined as binary, corresponding to the CRITICAL sample in which the word-based model outputs a different hypothesis from the morpheme sequence generated by the morpheme-based model.

$$y_i = \begin{cases} +1 \text{ if CRITICAL case is true} \\ -1 \text{ otherwise} \end{cases} \quad \textbf{(3-9)}$$

Note that the above judgment does not refer the correct hypotheses for the unsupervised training in which the training samples included the CRITICAL samples and the other samples in which both layers are unmatched but assumed as CRITICAL sample.

4) Lexicon design

The bigram feature is then generalized to all morpheme sequences in the text corpus prepared for language model training. If the evaluated bigram morphemes are classified to be merged, then this bigram sub-word granule is included in the lexicon. Otherwise they are left as separated morpheme units. Specifically, we try to search for sub-word entries that satisfy the evaluation. The search is exhaustively done from the beginning of all words by concatenating the following morphemes while the above-mentioned condition is met. If the condition is not met, the search is re-started there.

4. ASR results for segmented and concatenated lexica

All the segmented and concatenated lexica are separately applied to our Uyghur LVCSR system under a same AM. WER and lexicon size are compared.

4.1. Acoustic model construction

A speech corpus of general topics is prepared to construct an AM for Uyghur language. This AM is used for all the optimization experiments in this study. A held-out

test data set is prepared independently from readings of newspaper articles. Specifications of the data sets are summarized in Table 3.

Table 3 Statistics of speech corpus

corpus	sentences	speakers	total utterances	word tokens	time (hours)
training	13.7K	353	62K	895.1K	158.6
test	550	23	1468	14.7K	2.4

An acoustic model based on tri-phone HMMs with 3000 shared states and 16 Gaussian mixtures was trained for 34 Uyghur phones (8 vowels, 24 consonants, and 2 silence models). The acoustic features consist of 12 MFCCs, ΔMFCCs and ΔΔMFCCs together with Δpower and ΔΔpower.

Recent trends of using DNN achieved high accuracy in LVCSR experiments, especially for the under-resource languages. DNN can provide better contextual dependences between acoustic units, while the HMM only depends on the previous 1~2 units. In this study we used the RBM (Restricted Boltzmann Machines) based DNN architecture with 4 hidden layers and 1200 hidden units per layer to train a phoneme based Uyghur acoustic model based on the same speech data [29].

4.2. Lexical model construction

Lack of resource is one of the biggest problems for Uyghur natural language processing. It is difficult to have a large qualified corpus from a unique source (e.g. newspaper). So we selected texts from various publications like novels, newspapers, educational materials (history, science...). And we prepared a raw corpus of about 630k sentences which are from general topics. This corpus is prepared by removing all duplicated sentences, since it was a collection of different sources and may contain several copies of a same content. This text corpus separately segmented to word, morpheme, and pseudo-morpheme sequences, and n-gram language models are constructed separately based on each of them. Kneser-Ney smoothing is adopted. Vocabulary size, coverage and perplexity are calculated for each model. And, we keep the surface forms of morphemes same as in the words while they are split or merged, thus the words can be recovered simply by re-merging morphemes without any changes.

Two additional parameters which affect ASR system performances are investigated. One is the n-gram dimension; another is the cutoff-rate. Since the morpheme-based model is benefited from a much smaller vocabulary size, various n-gram dimensions are investigated for them in order to find out best ASR baseline results. The cutoff threshold also controls the lexicon size and ASR performance. Cutoff-F means that units with frequency less than F times are disregarded and treated as unknown.

4.3. ASR results on segmented lexica

Language models based on word, morpheme, and pseudo-morpheme units are separately constructed using the training text corpus, and the ASR results are compared as in Table 4. We can see that the pseudo-morpheme model outperformed other models, as it is already an optimized lexicon. However, the pseudo-morphemes are not fixed units, whose units and lexicon size change with a different training corpus, and exhibit a similar statistical property (perplexity) like words. Word-based model also outperforms the morpheme-based models with a larger lexicon size. When we have a smaller OOV for both sub-word units, the linguistic morpheme units may suffer from morpho-phonetic confusions. However, the morpheme-based model can be expanded to a huge vocabulary while the vocabulary of the word-based model is limited to the vocabulary of the training

corpus. Moreover, linguistic morpheme provides syntactic and semantic information which facilitates feature-based ASR and NLP.

Table 4 **ASR results for different baseline units (cutoff-2)**

LM names	WER (%)	lexicon size	word perplexity	OOV rate
word 3-gram	25.72	227.9k	2356	2.8%
word 4-gram	25.93	227.9k	1734	2.8%
morph 3-gram	28.96	55.2k	1733	0.3%
morph 4-gram	27.92	55.2k	1244	0.3%
morph 5-gram	29.31	55.2k	1144	0.3%
pseudo-morpheme 4-garam	25.04	133.4k	2314	0.8%

The Two-layer-lexica based concatenation approach utilizes both speech and text, while all other approaches are based on only text. So the two layers of lexica based baseline ASR results of a large speech data is necessary for the extraction of aligned CRITICAL samples. Table 5 shows the ASR performances of various models. The best n-gram dimensions with cutoff-2 parameter are selected as the baseline models. We can see from the results that word based model produce more correct hypothesis than the morpheme based model. Furthermore, we also investigated the effect of cutoff-rate to ASR performance of different lexica. We can see that pseudo-morpheme model produced the best ASR result. It is least affected by the cutoff-rates, while the word model is most vulnerable.

Table 5 ASR results for various baseline units

Baseline models		WER (%)	lexicon size	OOV
morpheme 4-gram	**cutoff-2**	**27.92**	**55.2k**	0.3%
	cutoff-5	28.11	27.4k	0.7%
word 3-gram	**cutoff-2**	**25.72**	**227.9k**	2.8%
	cutoff-5	26.64	108.1k	4.4%
pseudo-morph. 4-gram	**cutoff-2**	**25.04**	**133.4k**	**0.8%**
	cutoff-5	**25.01**	**94.5k**	**0.9%**

4.4. ASR results on concatenated lexica

The morpheme based text training data are transformed into various concatenated granules based text corpora and n-gram LMs separately built on them in order to evaluate their ASR performances.

4.4.1. Results of data-driven concatenation methods: Data-driven methods concatenate morphemes by various criteria. The concatenation can be made even across word boundaries in frequent unit sequence and mutual probability methods. The tuning of the threshold values for these methods are not so straight-forward, depending on the task and data set. Table 6 shows the best ASR results obtained by fine tuning.

Table 6 Result of data-driven concatenation approaches

Data-driven approaches	WER (%)	Lexicon size	OOV rate
Stem & word-endings	28.13	82.6K	0.5%
Frequent unit sequence	26.63	50.7K	0.8%
Mutual probability	25.60	53.3K	0.9%

4.4.2 Results of two-layer-lexica based approaches: The speech corpus used for acoustic model training is decoded by the two baseline models mentioned in section 4.3. ASR results are aligned and CRITICAL samples are extracted for the *error frequency* feature. WER and lexicon size are compared for 4-gram models with cutoff-5 pruning.

For error frequency feature, words misrepresented more than twice are extracted, and added to the morpheme lexicon. This approach can be iteratively applied to extract new candidates. WER and lexicon size within two iterations are listed in Table 7. On each iteration, new candidate words are added to the lexicon, until few candidates can be extracted. This simple method cannot include entries that are not included in the training CRITICAL samples.

Table 7 Results of word selection based on error frequency

Iterations	Baseline	First round	Second round
WER (%)	28.11	26.11	25.82
lexicon size	27.4K	40.4K	46.1K

4.4.3 Results on discriminative training: An SVM based discriminative evaluation function is trained on the aligned two-layer ASR results. Among the various features, bigram feature can efficiently capture the context compared to unigram and length features, so the redundant features are ignored in this paper. Extracted problematic bigram morphemes are concatenated, and generalized to the text training corpus. This method is more effective when conducted thoroughly in the sub-word level than the word level [23]. So the below discussions are sub-word optimization approaches.

First we investigate the supervised training by comparing with the correct transcription which means only CRITICAL samples (28.5%) are used for this training. Every bigram morpheme sequences are evaluated iteratively within word boundary. Then, it is propagated to the text training corpus. ASR result of this sub-word lexicon is shown in Table 8.

Table 8 supervised discriminative training results with bigram feature

method	WER (%)	lexicon size
SVM (cutoff-5)	25.42	45.1K

Now we extract all cases in which two layers are unmatched in the two-layer ASR hypothesis. Thus the training sample can include the 28.5% CRITICAL cases in which word-based model gives correct hypotheses while the morpheme-based model does not, and the majority of 68% cases in which both layers are misrecognized and not benefited by the supervised learning. So we can use an un-transcribed speech data to train this discriminative model. The dimension of the unigram features is 17K and that of the bigram is 53K in our speech training data. In the unsupervised experiment, the majority samples are defaulted as desired samples. Results in Table 9 shows that the unsupervised model outperformed the supervised method with the SVM learning method.

Table 9 unsupervised discriminative training results compared with baseline models.

Models		WER (%)	lexicon size	OOV
baseline morpheme		27.92	55.2K	0.3%
baseline word		25.72	227.9K	2.7%
SVM	cutoff-2	24.64	101.2K	0.7%

	cutoff-5	**24.61**	**55.1K**	0.9%

The result in Table 9 shows that the sub-word-based model trained with the bigram feature outperforms the best word based model in accuracy with the lexicon size of one fourth. And it is the most stable model, for the WER is not affected by different cutoff-rates.

4.4.4 Result on a statistical model based concatenation method: We can directly optimize from the morpheme sequence by using a proper algorithm for the concatenation approach. One direct and simple way is to adapt an existing tool to this task. We adapt the unsupervised word splitter, *Morfessor* tool, to concatenate *linguistic* morphemes. And we can simply do that by feeding morpheme based text to *Morfessor*. But the smallest granules now are the morphemes rather than characters. Thus the searching point of *Morfessor* algorithm is confined to morpheme boundaries only. The output is some kind of regrouping of morpheme particles.

We can see the result from Table 10 that both concatenated morphemes and pseudo-morphemes have an improved ASR performance which are comparable with the discriminative method, but with a larger vocabulary size.

Table 10 results of model based concatenation compared with other models

models	WER (%)	lexicon size	OOV
linguistic morpheme	28.11	27.4K	0.7%
pseudo-morpheme	25.01	94.5K	0.9%
model based morpheme concatenation	**24.96**	98.35K	0.9%

4.5. DNN based ASR results

The baseline word lexicon, morpheme lexicon, and the discriminatively optimized lexicon, which is obtained totally based on HMM based AM, are incorporated with the DNN based AM separately. And the ASR results are compared without changing the LMs and training corpora.

Table 11 DNN based comparison of various lexica

Models	**WER (%)**	**lexicon size**	**OOV**
baseline word	**16.50**	**227.9K**	**2.7%**
baseline morpheme	**14.50**	**55.2K**	**0.3%**
optimal lexicon	**12.89**	**55.1**	**0.9%**

Table 11 shows that under the DNN based AM, all lexical sets drastically improved the ASR accuracy with the same LMs as in the previous sections. Especially, we can see that the optimal lexicon that obtained from HMM based ASR results still outperformed other units in the DNN based model. This result demonstrates the generality of the two-layer-lexica based optimization approach.

5. Conclusions

This paper is a complete study on the lexicon design of Uyghur LVCSR system. Based on the derivational nature of the Uyghur language morphology, we have discussed particle segmentation approaches, and investigated statistical properties and ASR performances of segmented lexica. Some effective concatenation approaches for ASR systems are also investigated and results are compared. The optimized sub-word lexica proved to be better generalized than word or morpheme unit by exhibiting improved ASR performances. Pseudo-morpheme units, optimized directly from a raw text corpus, exhibits good ASR performance, but the extracted units are dependent on the training corpus.

Morpheme unit provides small lexicon, better statistical properties. It is a good foundation of concatenation optimization and convenient for downstream processing. The discriminative optimization approach outperformed all other methods, and generated a lexicon size within 64K (16 bit) which is impossible for the word units of languages which have derivational morphology. The optimized lexicon is very stable without having many susceptible parameters. And the unsupervised discriminative method is scalable for a large un-transcribed speech data.

Finally, this research provides a good example for the resource scarce languages which also have concatenative morphology. The results demonstrated the accuracy and generality of the optimization approaches as the optimal lexicon obtained by using the HMM based acoustic model also proved to be very effective when directly incorporated with the DNN based acoustic model.

Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC grant 61462085 and 61163032).

References

[1] M. Ablimit, G. Neubig, M. Mimura, S. Mori, T. Kawahara, A. Hamdulla, "Uyghur *Morpheme-based* Language Models and *ASR,"* In Proc. ICSP, Beijing, (**2010**).
[2] A. Lee, T. Kawahara, and K. Shikano, "Julius -- an open source real-time large vocabulary recognition engine," In Proc. Eurospeech, pp. 1691--1694, (**2001**).
[3] O.-W. Kwon and J. Park, "Korean large vocabulary continuous speech recognition with morpheme-based recognition units," Speech Communication, vol. 39, pp. 287–300, (**2003**).
[4] O-W. Kwon, "Performance of LVCSR with morpheme-based and syllable-based recognition units," In Proc. IEEE-ICASSP, pp.1567–1570, (**2000**).
[5] K. Hacioglu, B. Pellom, T. Ciloglu, O. Ozturk, M. Kurimo, M. Creutz, "On Lexicon Creation for Turkish LVCSR," In Proc. Eurospeech, (**2003**).
[6] E. Arisoy, D. Can, S. Parlak, H. Sak, M. Saraclar, "Turkish Broadcast News Transcription and Retrieval," IEEE Trans. Audio, Speech & Language Processing, vol. 17, no. 5, pp. 874-883, (**2009**).
[7] H. Sak, M. Saraclar, T. Gungor, "Morpholexical and Discriminative Language Models for Turkish Automatic Speech Recognition," IEEE Trans. Audio, Speech & Language Processing, vol. 20, no. 8, pp. 2341-2351, (**2012**).
[8] M. Larson, D. Willett, J. Kohler, G. Rigoll, "Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches," In Proc. Interspeech, (**2000**).
[9] M. Nuβbaum-Thom, A. El-Desoky Mousa, R. Schluter, Hermann Ney, "Compound Word Recombination for German LVCSR," In Proc. Interpeech, (**2011**).
[10] B. Xiang, K. Nguyen, L. Nguyen, R. Schwartz, J. Makhoul, "Morphological decomposition for Arabic broadcast news transcription", In IEEE-ICASSP (**2006**).
[11] A. El-Desoky, C. Gollan, D. Rybach, R. Schluter, H. Ney, "Investigating the use of morphological decomposition and diacrit", In Proc. Interspeech, (**2009**).
[12] M. Jongtaveesataporn, I. Thienlikit, C. Wutiwiwatchai, S. Furui, "Lexical units for Thai LVCSR," Speech Communication, pp.379–389, (**2009**).
[13] M. Creutz, "Introduction of the morphology of natural language: Unsupervised morpheme segmentation with application to automatic speech recognition", PhD. Thesis, Helsinki University of Technology, Finland, (**2006**).

[14] J. Goldsmith, "Unsupervised learning of the morphology of a natural language", Computational linguistics, vol. 27, June. (**2001**)
[15] B. Roark, M. Saraclar, and M. Collins, "Discriminative n-gram language modeling," Computer Speech and Language, 21(2):373–392, (**2007**).
[16] Graham Neubig, "Unsupervised Learning of Lexical Information for Language Processing Systems," PhD thesis, Kyoto University. (**2012**).
[17] M. Creutz, T. Hirsimaki, M. Kurimo, A. Puurula, J. Pylkkonen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraclar, and A. Stolcke. "Morph-based speech recognition and the modeling of out-of-vocabulary words across languages," ACM Trans., Speech & Language Processing, vol. 5, no. 1, pp. 1-29, (**2007**).
[18] T. Pellegrini, L. Lamel, "Using phonetic features in unsupervised word decompounding for ASR with application to a less-represented language," In Proc. Interspeech, (**2007**).
[19] Hirsimaki T. et al., "Unlimited vocabulary speech recognition with morphlanguage models applied to Finnish[J]," Computer Speech and Language, 20(4):515-541, (**2006**).
[20] P. Mihajlik, Z. Tuske, B. Tarján et al., "Improved Recognition of Spontaneous Hungarian Speech—Morphological and Acoustic Modeling Techniques for a Less Resourced Task," IEEE Transactions on Audio, Speech, and Language Processing,18(6)：1588-1600, (**2010**).
[21] P. Mihajlik, T. Fegyo, Z. Taske, P. Ircing, "A morpho-graphemic approach for the recognition of spontaneous speech in agglutinative languages-Like Hungarian," In Proc. Interspeech, pp. 1497-1500, (**2007**).
[22] E. Arisoy, M. Saraclar, B. Roark, I. Shafran, "Discriminative language modeling with linguistic and statistically derived features," IEEE Trans. Audio, Speech & Language Processing, vol. 20, no. 2, pp. 540-550, (**2012**).
[23] M. Ablimit, T. Kawahara, A. Hamdulla, "Lexicon optimization based on discriminative learning for automatic speech recognition of agglutinative language," Speech Communication, (**2014**).
[24] T. Shinozaki and S. Furui, "A New Lexicon Optimization Method for LVCSR Based on Linguistic and Acoustic Characteristics of Words", In Proc. ICSLP, pp. 717-720, (**2002**).
[25] G. Saon, M. Padmanabhan, "Data-Driven Approach to Designing Compound Words for Continuous Speech recognition," IEEE Trans. Speech and Audio Processing, Vol.9, No. 4, pp. 327-331, (**2001**).
[26] M. Ablimit, A. Hamdulla, T. Kawahara, "Morpheme Concatenation Approach in Language Modeling for Large-Vocabulary Uyghur Speech Recognition," In Proc. Oriental-COCOSDA Workshop, (**2011**).
[27] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," Journal of Machine Learning Research, 9: 1871-1874, (**2008**).
[28] T. Kawahara et al., "Free software toolkit for Japanese large vocabulary continuous speech recognition," In Proc. ICSLP, Vol.4, pp.476–479, (**2000**).
[29] Povey, Daniel and Ghoshal, Arnab and Boulianne, Gilles and Burget, Lukas and Glembek, Ondrej and Goel, Nagendra and Hannemann, Mirko and Motlicek, Petr and Qian, Yanmin and Schwarz, Petr and Silovsky, Jan and Stemmer, Georg and Vesely, Karel, "The Kaldi Speech Recognition Toolkit," IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, (**2011**).

Authors

MijitAblimit received his M.S. in 2011, Ph.D. in 2013, in Information Science and Engineering respectively from Xinjiang University of China and Kyoto University of Japan. Now, He is an associate professor in the School of Information Science and Engineering, Xinjiang University, and doing his research work in the Computer Science and Technology Postdoctoral Research Center of Xinjiang University. His research interests include language, speech processing, and pattern recognition.

Tatsuya Kawahara received B.E. in 1987, M.E. in 1989, and Ph.D. in 1995, all in information science, from Kyoto University, Kyoto, Japan. In 1990, he became a Research Associate in the Department of Information Science, Kyoto University. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a Professor in the Academic Center for Computing and Media Studies and an Affiliated Professor in the School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR and NICT. He has published more than 250 technical papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been conducting several speech-related projects in Japan including a free large vocabulary continuous speech recognition software project (http://julius.sourceforge.jp/) and the automatic transcription system for the Japanese Parliament (Diet). Dr. Kawahara received the 1997 Awaya Memorial Award from the Acoustical Society of Japan and the 2000 Sakai Memorial Award from the Information Processing Society of Japan. From 2003 to 2006, he was a member of IEEE SPS Speech Technical Committee. From 2011, he is a secretary of IEEE SPS Japan Chapter. He was a general chair of IEEE Automatic Speech Recognition & Understanding workshop (ASRU 2007). He also served as a tutorial chair of INTERSPEECH 2010. He is a senior member of IEEE.

Askar Hamdulla received B.E. in 1996, M.E. in 1999, and Ph.D. in 2003, all in Information Science and Engineering, from University of Electronic Science and Technology of China. In 2010, he was a visiting scholar at Center for Signal and Image Processing, Georgia Institute of Technology, GA, USA. Currently, he is a professor in the School of Software Engineering, Xinjiang University. He has published more than 140 technical papers on speech synthesis, natural language processing and image processing. He is a senior member of CCF and an affiliate member of IEEE.