Uyghur Morpheme-based Language Models and ASR

Mijit Ablimit[†], Graham Neubig[†], Masato Mimura[†],
Shinsuke Mori[†], Tatsuya Kawahara[†], Askar Hamdulla[††]

Uyghur language is an agglutinative language in which words are formed by suffixes attaching to a stem (or root). Because of the explosive nature in vocabulary of the agglutinative languages, several morpheme-based language models are built and experiments are implemented. Morpheme is the smallest meaning bearing unit. In this research, morpheme is referred to any of prefix, stem, or suffix. As a result, a large vocabulary ASR system is built on the basis of Julius system. Several ASR results on language models based on different units (word, morpheme, and syllable) are compared.

ウイグル語の形態素に基づく言語モデルと音声認識システム

アブリミテ・ミジテ[†]，　ニュービッグ・グラム[†]，
三村正人[†]，森信介[†]，河原達也[†]，ハムヅラ・アスカ[††]

ウイグル語は膠着性の言語で、単語は語幹(語根)に接尾辞が付くことによって構成される。膠着性の言語の語彙は爆発的に大きくなるので、形態素に基づく言語モデルを構成し、比較実験を行なった。形態素は意味を持つ最小の単位であり、本研究では接頭辞・語幹・接尾辞のいずれかを指す。Julius を用いて大語彙連続音声認識システムを構成し、異なる単位(単語・形態素・音節)の言語モデルの比較を行なった。

1. Uyghur language and morphological units

Uyghur belongs to the Turkish language family of the Altaic language system. At present, Uyghur is written in Arabic scripts with some modifications. There are 32 phonemes in Uyghur, 8 vowels and 24 consonants; one phoneme is recorded by one character. Sentences in Uyghur consist of words, which are separated by space or punctuation marks. Uyghur words consist of some morphological units without any splitter between them.

(Example.1 morpheme and syllable segmentation)

Müshükning kəlginini korgən chashqan hoduqup qachti.
(*The mouse escaped by the sight of cat.*)
Müshük+ning kəlgən+i+ni kor+gən chashqan hoduq+up qach+ti. (morpheme sequence)
Mü+shük+ning kəl+gi+ni+ni kor+gən chash+qan ho+du+qup qach+ti. (syllable sequence)

The morpheme structure of Uyghur word is ***“ prefix + stem + suffix1 + suffix2 + … ”***. A root (or stem) is attached in the rear by zero to many (longest is about 10 suffixes or more) suffixes. A few words can be added with a prefix (only one) in the head of a stem, and only 7 (difficult to find more) prefixes are used in this research. 108 suffix types are defined and collected, according to their semantic and syntactic functions, which can be extracted to 305 surface forms. The surface realizations of the morphological structure are constrained and modified by a number of language phenomenon such as insertion, deletion, phonetic harmony, and disharmony (vowel assimilation, vowel weakening [1][2]). Suffixes that make semantic changes to a root are derivational suffixes. Suffixes that make syntactic changes to a root are inflectional suffixes. A root linked with the derivational suffixes becomes a stem. So the root set is included in the stem set. Sometimes the words “stem” and “root” are used without distinguishing. To keep the versatile nature of language, we keep different segmentation forms of a same word in our training corpus.

(Example.2 different morpheme segmentation of the same word)

oqutquchi (teacher{stem}) = oqut(teach){root} + quchi(er) {suffix}
yazghuchi = yaz(write)+ghuchi(er)
hesablinidu = hesab+la+n+idu,　hesab+lan+idu;

[†] 京都大学情報学研究科
Kyoto University, School of Informatics
[††] 新疆大学信息学院
Xinjiang University, Information Institute

Syllables in Uyghur language is regular, and the general format is "CV[CC]" (C stands for consonant, V stands for vowel)[1]. Because of the direct importing of foreign words, new syllable formats are added such as "CCV[CC]" from some European languages, and "CVV[C]" from Chinese.

2. Segmentation of morphological units

2.1 Morpheme segmentation

An Uyghur morpheme segmenter has been developed by using statistical methods. In our segmentation, our primary goal is to catch the different forms of stem, not root. This will expand the size of stem vocabulary, but is more convenient for analyzing semantic and syntactic context of words.

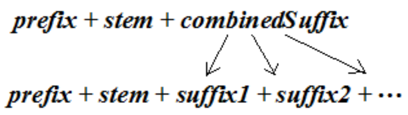
[Corpus preparation] A text corpus of 10025 sentences and their manual segmentations are prepared. These sentences are collected from general topics, unrelated. More than 30K stems are prepared independently and used for the segmentation task.

Table1. manually segmented morpheme corpus

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

[Method] For a candidate word, all the possible segmentation results are extracted in reference for both stem and suffix, and their probabilities are computed to get the best result.

At first, a word is split into two parts, a stem and a combined suffix, and several possible stem-suffix pairs are obtained.


prefix + stem + combinedSuffix
prefix + stem + suffix1 + suffix2 + ···

Then, the suffix is segmented into singular-suffixes, because each combined suffix (or stem endings in some papers) may have several different singular-suffix segmentations.

There are several problems in the segmentation. First, assimilation [1][2] (weakening or disharmony in some papers) should be recovered to standard surface forms. Second is the morphological change, which is deletion and insertion. Third is the phonetic harmony [2] which causes different surface forms of a same suffix. Fourth is the ambiguity (there are many reasons for this).

(Example.3 problems in morpheme segmentation)

(1) almini = alma + ni , almiliring = alma + lar + ing (weakening) ;
(2) oghli = oghul + i , kaspi = kasip + i (deletion) ;
(3) qalmaytti = qal+may+[t]+ti , binaying = bina+[y]+ing ; (insertion) ;
(4) yurttin = yurt+ tin ; watandin =watan + din (phonetic harmony) ;
(5) hesablinidu= hesab+la+n+idu = hesab+lan+idu; berish=bar(go/have)+ish, berish= bər(give)+ish; (ambiguity)

Generally, an intra-word bi-gram method based on the following probabilities is used, and the identification of stem-suffix boundary is the most important part in segmentation,

$$\begin{cases} \mathbf{P(stem,\ firstSuffix)} \\ \mathbf{P'(stem)P(anySuffix\,|\,stem)} & \text{for smoothing} \end{cases}$$

in which

$$\mathbf{P'(stem)} = \frac{\mathbf{stemFrequency}}{\mathbf{(stemToken{+}stemVocabulary)}}$$

$\mathbf{P(anySuffix\,|\,stem)}$ probability of a stem linked with a suffix

For insertion, we add the inserted phoneme to the subsequent suffix, and form a new surface form of the same suffix type. For deletion, because it happens in the stem only, a list of deleted stems are learned from the training corpus.

[Results] We split the corpus to the training corpus of 9025 sentences, and the test corpus of 1000 sentences. Word coverage is 86.85%. Morpheme coverage is 98.44%. The morpheme segmentation accuracy is **97.66%** which is the percentage of the exact match of all morphemes in automatic segmentation compared with manual segmentation. Generally two kinds of ambiguity exist in our segmentation. One is because of the definition of the stem set, the other is because of the sound harmony.

(Example.4 ambiguity during morpheme segmentation)

1.oqut(teach) , oqutquchi(teacher)
2.ish(job) ishlə (do), ishləp(done), ishləpchiqirix (produce)
3.berish=bar(go/have)+ish, berish = bər(give)+ish

In the first and second examples, several stems come out from one root. As we can see from this example, stem may be more convenient for practical applications than root. And the flexibility in segmentation should also reflect the flexibility of language itself. So we keep different segmentations of a same word in our learning corpus. However, this segmentation tool has only one segmentation result for a candidate word. Flexible segmentation needs more context analysis.

In the third example, the weakened stem (bar or bər) has a same surface form when attached by some suffixes. Both words are frequent words, and both results have high probabilities, but only the most probable one is produced in our tool.

2.2 Syllable segmentation

Syllable is another clear morphological unit in Uyghur language. The Uyghur words in general CV[CC] syllable format consist of about 99.1% of all words in our corpus. The words in the format of foreign syllables are about 0.6%. Except the misspelled words (around 0.3% by estimation), all words can be correctly segmented with our rule-based syllable segmenter. There may be ambiguities with a few words which are in the foreign syllable format. There are no changes in surface forms after syllable segmentation

3. Tri-gram language models (LM) on different units

3.1 Language models of different units

Lack of resource is one of the biggest problems for Uyghur language processing. From various publications, we prepared a raw corpus of about 630k sentences which are from general topics like novels, newspapers, books (history, science...). This corpus is prepared by removing all duplicated sentences, as it was a collection of different sources and may have many copies of same content. We segmented this corpus separately to morphemes and syllables, and built three tri-gram language models based on three different units: word, morpheme and syllable.

Words in Uyghur sentences are naturally separated by space or punctuation marks. All punctuation marks are removed in following experiments to keep the coverage and perplexity consistent in the LM experiment and ASR experiment.

Changes in the surface forms, especially the assimilation, cause problems for practical applications of morpheme based LMs. In Uyghur language, speech is recorded as pronounced. When a word is segmented, if there is assimilation, usually it is recovered to the standard surface format. We keep the surface forms of morphemes same as in the words, thus the words can be recovered simply by connecting morphemes without any changes.

Example.5 changes in surface forms:

teghi = tagh+i(recovered) ;
teghi = tegh+i(keep as in words) ;
almiliringiz = alma+lar+i+ngiz(recovered) ;
almiliringiz = almi+lir+i+ngiz(keep as in words)

These may cause some ambiguity in morphemes, but does not degrade segmentation accuracy. Without changing the surface forms of morphemes, we conducted tri-gram language model experiments.

In order to preserve the word boundary information, we either add a symbol for a word boundary between syllables and characters, or label the position of a morpheme. Among the units, only morpheme is the meaning bearing unit. Syllables and Characters are relatively random sequences. For syllable and character units, a word boundary symbol is added between syllables or characters in the place of word boundary. For morphemes, the prefix and suffix are labeled, nothing added to stem. This is for recovering the words from morphemes by simply connecting them together.

Exp.6 inserting word boundary in units:
Kishilər wəqədin bihəwər qaldi.
Kishi _lər wəqə _din bi_ həwər _qaldi.(morpheme)
Ki+shi+lər_wə+qə+din_bi+hə+wər_qal+di.(syllable)

Tri-gram models are built on word, morpheme, and syllable units, respectively; Kneser-Ney smoothing is adopted. Unknown word model is used, and words appeared only once are considered as unknown. Coverage and perplexity are calculated for each model.

As a test corpus, 11888 sentences are held out with the character size of 1460.8k, Table 2 shows statistics of the test corpus. From the statistics, a word unit is segmented into about two morphemes and three syllables on average. The remaining 620K sentences are used as a training corpus. Fig.1-4 and Table 3 show the results. The result shows that the morpheme-based language model performs comparably to the word-based language model with a much smaller size.

Table 2. statistics of test corpus

units	word	morph	syllable
tokens	217k	408.64k	592.57k
vocabulary	47k	15.34k	3.64k

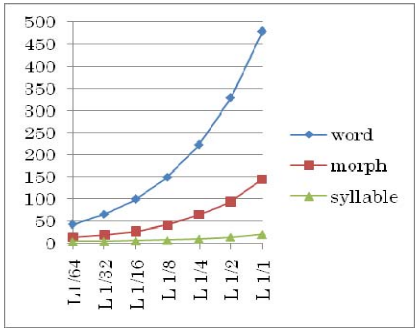


Fig.1 vocabulary size of different units

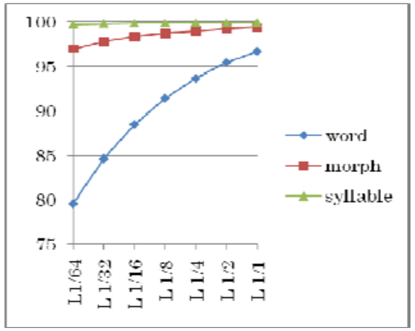


Fig.2 uni-gram coverage (%) of different units

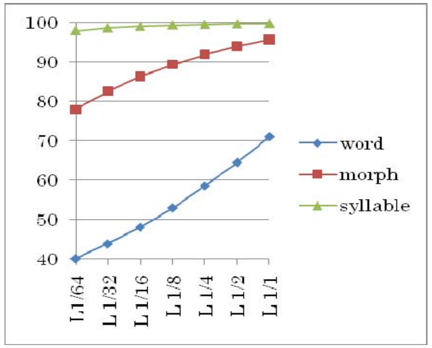


Fig.3 bi-gram coverage (%) of different units

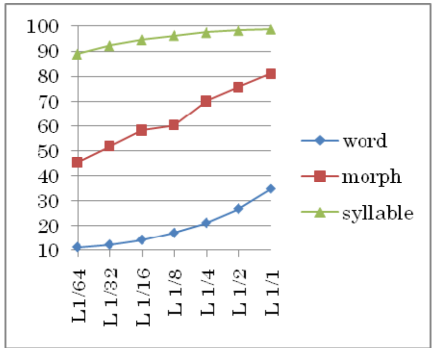


Fig.4 tri-gram coverage (%) of different units

Table 3. perplexity by tri-gram models of different units

training corpus	perplexity			perplexity normalized by words	
	word	morph	syllable	morph	syllable
L1/64	23566	162.6	16.1	14384	27740
L 1/32	14376	126.8	14.9	8987	20919
L 1/16	9153	103.3	14.1	6119	17037
L 1/8	5935	86.1	13.5	4343	14640
L 1/4	3847	73.6	13.2	3232	13148
L 1/2	2416	63.5	12.9	2447	12078
L 1/1	1408	54.8	12.6	1860	11335

3.2 Comparison of different n-grams

Then, we compare n-gram models of different lengths. Because of the memory limitation, we can only calculate until 5-gram for word and morpheme units, 6-gram for syllable unit, and 10-gram for character unit. To compare the results, the perplexity is normalized in reference to the word unit. Table 4 shows the result.

The morpheme and syllable models are significantly improved with longer n-grams, and the morpheme-based model performs better than the word-based model.

Table 4. normalized perplexity of n-gram models of different units

unit	word	morph	syllable	char
1-gram	21321	427628	110014618.	30014487856
2-gram	2210	5651	168482	140025078
3-gram	1408	1866	11337	4498647
4-gram	1260	1183	3349	217874
5-gram	1234	985	1901	29051
6-gram			1425	9186
7-gram				4743
8-gram				3113
9-gram				2397
10-gram				2032

4. Uyghur speech recognition system

We also built an ASR system using the language models, on the basis of Julius system. Julius is open-source large-vocabulary continuous speech recognition (LVCSR) software for researchers and developers. The acoustic models and language models are easily pluggable, and you can build various kinds of speech recognition systems by preparing your own models suitable for the task. It also adopts standard formats to handle other toolkits such as HTK, CMU-Cam SLM toolkit, etc.

4.1 Uyghur acoustic model

A relatively large speech corpus was prepared to build an acoustic model of Uyghur.

[Training corpus] Total 62K utterances are recorded with about 13.7K different sentences, spoken by 353 persons. These sentences are collected from general topics. The speech signals are sampled at 16 kHz with a resolution of 16 bits.

[Test corpus] 550 sentences from the news corpus are used for a test corpus; each sentence is read by at least one male and one female, total 23 people. As a result, 1248 utterances are used.

There are 32 phonemes in Uyghur, 8 vowels and 24 consonants. One character corresponds to one phoneme, so there are 32 different characters, with one additional character which is actually a syllable segmentation mark. We used 34 basic phonemes including silence. HTK is used to build three-state HMM with 16-Gaussian mixture models. A standard 38-dimensional feature vector is used.

4.2 Uyghur ASR experiments on different units

For the vocabulary file of the ASR, we did spell checking by some morphological analysis, such as syllable format and word format. So the vocabulary gets relatively smaller, and this also improves the ASR accuracy.

The beam size in all ASR experiments is 10,000. Because of the huge vocabulary of the word-based language model, a large beam size is used in decoding.

Five different language models are built using the training corpus, and ASR results are compared. The word boundary symbol is added to all units other than word unit.

①Word-based language model.

②Morpheme-based language model.

③FMS (Frequent Morpheme Sequence) based language model. FMS unit is built by combining morpheme sequences of frequency of at least 500 times in the training corpus.

④Stem-Suffix (stem endings, or word endings) based ASR; the word is segmented into two parts: stem and combined suffix. In other words, all the singular suffixes are combined. Singular suffixes are relatively shorter units, and they are the frequent sequence.

⑤Syllable based language model.

As we can see, except the word and syllable-based LMs, other three types of LMs are based on combinations of morphemes. The units other than word unit are recovered to words. Because the word boundary is preserved, the morphemes can be recovered to words by simply connecting them. For the morpheme unit, we conduct additional ASR experiments using 4-gram and 5-gram language models. The results are shown in Table 5

The vocabulary of syllable-based ASR is 6.58k and the syllable error rate is 28.73%. Word boundary is not taken into consideration for syllable.

The results show that the word-based language model performs best. 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, morpheme provides syntactic and semantic information which facilitates feature-based ASR and NLP.

Table.5 ASR error rates for different LMs

LM names	Words	FMS-500	Stem-Suffix	morph-3gram	morph-4gram	morph-5gram
vocabulary	227.9k	274.97k	74.5k	55.2k	55.2k	55.2k
Morpheme Error Rate (%)	18.88	21.28	21.69	22.73	21.64	22.98
Word Error Rate (%)	25.58	28.14	28.13	28.96	27.92	29.31

5. Conclusion

During the design and implementation of the morpheme segmenter, we manually segmented and standardized the Uyghur morphemes, especially the suffixes. By collecting large text and speech corpora, we have obtained a reliable statistics for Uyghur language on three different units. We also built an ASR system based on a variety of language models. In the ASR evaluations, word-based model performed best, like Turkish [5], but we expect the morpheme-based language model paved us a huge road for the future development of Uyghur language processing.

References

1) Gulila Adongbieke, Mijiti Abulimiti. Research on Uighur Word Segmentation, 2004.11, Journal of Chinese information Processing,

2) M.Ablimit, M.Eli, and T.Kawahara "Partly supervisedUyghur morpheme segmentation" In Proc.Oriental-COCOSDA Workshop, 2008, pp.71—76

3) Batuer Aisha, Maosong Sun. A Statistical Method for Uyghur Tokenization. Proceedings of IEEE International Conference on Natural Language. 2009.

4) Askar Hamdulla, Dilmurat Tursun. An Acoustic Parametric Database for Uyghur Language. Proceedings of the 2009 International Joint Conference on Artificial Intelligence.

5) Ebru Arisoy , Helin Dutağaci , Levent M. Arslan, A unified language model for large vocabulary continuous speech recognition of Turkish, Signal Processing, v.86 n.10, p.2844-2862, October 2006

6) Afify and Ruhi Sarikaya. On the Use of Morphological Analysis for Dialectal Arabic. in proceedings INTERSP2006-70, Speech Recognition

7) Ruhi Sarikaya and Mohamed Afify and Yuqing Gao. JOINT MORPHOLOGICAL-LEXICAL LANGUAGE MODELING (JMLLM) FOR ARABIC. In proceedings ICASSP 2007-4-1031,

8) Hasim Sak, Murat Saraclar and Tunga Gungor. Intergrating Morphology into Automatic Speech Recognition. In proceedings. SAK:Turkish. 2009

9) Hasim Sak, Murat Saraclar and Tunga Gungor. Morphology-based and Sub-word Language Modeling for Turkish Speech Recognition. Proceedings SAK:Turkish. 2010,

10) 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.

11) B.Roark and M.Saraclar and M.Collins. Discriminative N-gram Language Modeling. ROA: discriminative-LM. CSL. 2007