International Journal of Future Generation Communication and Networking
Vol. 9, No. 2 (2016), pp. 59-72
http://dx.doi.org/10.14257/ijfgcn.2016.9.2.07

Stem-Affix based Uyghur Morphological Analyzer

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

[1]*Institute of Information Science and Engineering, Xinjiang University, China*
mijit601@gmail.com
[2]*School of Informatics, Kyoto University, Kyoto, Japan*
kawahara@ar.media.kyoto-u.ac.jp
[3]*School of Software, Xinjiang University, Urumqi, China*
askarhamdulla@gmail.com

Abstract

Uyghur language is an agglutinative language in which words are derived from stems (or roots) by concatenating suffixes. This property makes a large number of combinations of morphemes, and greatly increases the word-vocabulary size, causing out-of-vocabulary (OOV) and data sparseness problems for statistical models. So words are split into certain sub-word units and applied to text and speech processing applications. Proper sub-word units not only provide high coverage and smaller lexicon size, but also provide semantic and syntactic information which is necessary for downstream applications. This paper discusses a general purpose morphological analyzer tool which can split a text of words into sequence of morphemes or syllables. Uyghur morpheme segmentation is a basic part of the comprehensive effort of the Uyghur language corpus compilation. As there are no delimiters for sub-word units, a supervised method, combined with certain rules and a statistical learning algorithm, is applied for morpheme segmentation. For phonetic units like syllable and phonemes, pure rule-based methods can extract with high accuracy. Most common and proper sub-words for various applications can be the linguistic morphemes for they provide linguistic information, high coverage, low lexicon size, and easily be restored to words. As the Uyghur language is written as pronounced, phonetic alterations of speech are openly expressed in text. This property makes many surface forms for a particular morpheme. A general purpose morphological analyzer must be able to analyze and export in both standard and surface forms. So the morpho-phonetic alterations like phonetic harmony, weakening, and morphological changes are summarized and learnt from training corpus. And a statistical model based morpheme segmentation tool is trained on the corpus of aligned word-morpheme sequences, and applied to predict possible morpheme sequences. For an open test set, with word coverage of 86.8% and morpheme coverage of 98.4%, the morpheme segmentation accuracy is 97.6%. This morpheme segmentation tool can output both on the standard forms and on the surface forms without costing segmentation accuracy. Furthermore, for various basic lexical units of word, morpheme, and syllable, the statistical properties are compared as a comprehensive effort of the Uyghur language corpus compilation.

Keywords*: Uyghur; morpheme; morphology, phonetics, vowel weakening*

1. Uyghur Language and Morphological Structure

Uyghur belongs to the Turkish language family of the Altaic language system. Uyghur text is written as pronounced, each phoneme is recorded by a character, total 32 characters for 32 phonemes (8 vowels and 24 consonants). Sentences in Uyghur consist of words which are separated by space or punctuation marks. The words consist of smaller sub-word morphological units like morphemes or phonetic

ISSN: 2233-7857 IJFGCN

units like syllables and phonemes. Table 1 shows an example of various morphological units of words, morphemes, and syllables.

Table 1. Example of Morphological Units of Uyghur Language

Uyghur sentence	Müshükning	kəlginini	korgən	chashqan	hoduqup	qachti.
morpheme sequences	Müshük+ning	kəl+gən+i+ni	kor+gən	chashqan	hoduq+up	qach+ti
syllable sequences	Mü+shük+ning	kəl+gi+ni+ni	kor+gən	chash+qan	ho+du+qup	qach+ti
in English	The mouse seeing the cat coming was startled and escaped.					

Uyghur language has a strong syllable structure, CV[CC] (plus a few imported foreign syllable structures)[1], each syllable is centered by only one vowel, thus has a stable phonetic structure. Rule based segmentation can extract phonetic units like phonemes and syllables with high accuracy [2-3].

The morpheme here is defined as the smallest functional unit. The morpheme structure of an Uyghur word is "prefix + stem + suffix1 + suffix2 + …". A stem (or root) is followed by zero to many suffixes. A few words have a prefix (only one) in the head of a stem. Suffixes are defined and collected, according to their semantic and syntactic functions. Generally, there are two types of suffixes, derivational suffixes that make semantic changes, and inflectional suffixes that make syntactic changes. When a root is followed by a derivational suffix, it becomes a stem. Here the root set is included in the stem set. Some examples are shown in Table 2.

Table 2. Examples of Morpheme Segmentation to Stems or Roots

stems	root+ suffixes
oqutquchi (teacher)	oqut (teach) + quchi (er) {suffix}
yazghuchi (writer)	yaz (write)+ghuchi (er)
hesablinish (calculate, consider)	hesab (calculus)+la+n+idu, hesab+lan+idu

The surface realizations of the morphological structure are constrained and modified by a number of language phenomenon such as insertion, deletion, phonetic harmony and weakening (or disharmony, assimilation [4-6]). The morphemes have their standard forms and surface forms. Certain suffixes have 4 different surface forms resulted from phonetic harmony, and more surface forms are resulted from morphological changes and weakening. Mainly the morpho-phonetic changes are the result of the strong syllable bond in Uyghur language. A general purpose morphological analyzer tool must consider morpho-phonetic changes of sub-word units and handle both standard and surface forms, thus can be applied to different research purposes.

Morphemes can be segmented in a supervised manner in order to have higher segmentation accuracy. There are several ways of inducing morphemes. The rule based methods utilize some linguistic knowledge such as morpho-phonetic rules, and dictionaries like a stem list and a suffix list. The morpho-phonetic variations can be learnt by comparing morphemes with their aligned word sequences. For statistical model based approaches, a statistical learning model can be constructed and trained on a manually prepared training corpus to extract most probable morpheme sequences. Furthermore, there are also unsupervised segmentation approaches which split words into morpheme-like units from a raw text corpus

without considering linguistic properties [6-8]. In this paper, we focus on morphemes which are strictly meaning bearing units.

2. Inducing Morphological Units

As the Uyghur language has a stable syllable structure, a rule-based method can segment phonetic units, phonemes and syllables with high accuracy [1][10]. Our main target on this research is morpheme segmentation which relies on morpho-phonetic analysis.

A stem-centered approach is applied for the supervised morpheme segmentation in this research. Stems in Uyghur language remain fairly unchanged after suffixation compared to the suffixes which heavily suffer from morpho-phonetic changes, and easily being confused. We can manually summarize and classify the phonetic rules, and implement them for the rule-based morpheme segmentation task. This rule-based method utilizes a stem list and a combined suffix (word-ending) list. But the words which are not covered by these two lists cannot be correctly split. But, a statistical learning algorithm has a flexibility to learn probable morpheme sequence from a training corpus, and predict OOV words [11-16]. However, both rule-based and statistical modeling methods need an analyzer of morpho-phonetic changes like syllable rules, phonetic harmony, weakening, deletion, insertion, and substitution.

2.1. Phonetic Rules in Uyghur Language

When the morphemes are concatenated, the surface forms often change (or harmonize) around the boundary according to certain phonetic rules. The phonetic rules are classified and corresponding samples are extracted and learnt from training corpus in order to incorporate in the morpheme segmentation task.

2.1.1 Syllable Structure

Syllable is a clear phonetic unit in Uyghur language. The conventional syllable structure is CV[CC]. Generally, the words in this format consist of about 99.1% of all words. The words in other syllable formats, which are imported foreign syllables, are about 0.6%. There are misspelled words of about 0.3% in our corpus. The rule-based segmenter can correctly segment words into syllables with 99.5% accuracy. There may be ambiguities for a few foreign syllables, but, there are no changes in the surface forms after syllable segmentation.

2.1.2 Phonetic Harmony

Phonetic harmony is the harmony of vowels and consonants in the interface of the morpheme boundaries according to their accent. Phonetic harmony is the basic controlling rule in the root-suffix linkage. There are two types of phonetic harmony in Uyghur language: consonant harmony and vowel harmony. When certain morphemes are concatenated, the last vowel of previous morpheme is harmonized with the first vowel of next morpheme according to their tongue position, this phenomena is called vowel harmony. Similarly the last consonant of previous morpheme is harmonized with the first consonant of next morpheme according to their surdness, this phenomena is called final consonant harmony. Phonetic harmony is a complex phenomenon which is not fully predicted according to rules. There are four types of harmonization which caused different surface forms of morphemes as shown in Figure 1. Usually stems are suffered from vowel weakening, and suffixes have both phonetic harmony and weakening.

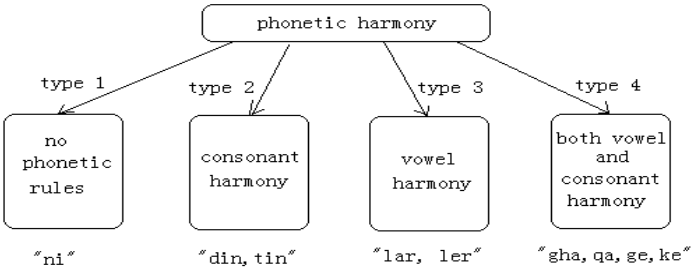


Figure 1. Phonetic Harmony in Uyghur Language

Type 1: This kind of suffix has only one fixed form. For example, "ni, ning", for the stem "adəm" (person).

"adəmning=adəm+ning (correct)"

"adəmni=adəm+ni (correct)"

Type 2: Consonants at the interface of morphemes must be harmonized according to the phoneme type of surd or sonant. For example, for the suffix type "din, tin"

"adəmdin=adəm+din (correct)"

"adəmtin=adəm+tin (wrong)"

Type 3: Vowels at the interface of morphemes must be harmonized according to the articulation point of vowels. For example, for the suffix type "lar, lər"

"adəmlar=adəm+lər (correct)"

"adəmlər=adəm+lar (wrong)"

Type 4: In this type, morphemes are harmonized according to both of type2 and type3, for example, for the suffix type "gha, qa, gə, kə"

"adəmge=adəm+gə (correct)"

"adəmgha=adəm+gha (wrong)"

"adəmqa=adəm+qa (wrong)"

"adəmke=adəm+kə (wrong)"

2.1.3 Vowel Weakening

Vowel weakening is another common phonetic change which is reflected in written text and cannot be directly predicted. The possibly weakened syllable inside a word must be recovered in order to match it with the stem list and suffix list. As we do not know which syllable is weakened, our method is to check one by one by recovering certain vowels. A candidate word is segmented to syllables to find possibly weakened phonemes "i" and "e", and recover them separately to "a" and "ə". Then recovered words are tested by matching with the stem list. Several different segmentation results may be obtained by over-segmenting to a shorter stem and causing ambiguity. For example, the word "almisi (somebody's apple)" can be segmented to three different results.

"almisi = alma + si ",

"almisi = al (take) + ma + si ",

"almisi = almas (diamond) + i."

In these examples, first and third are correct segmentations with different meaning and the second one are an incompatible combination of stem and word-ending. Only semantic or context analysis can determine the best result. In the meantime, the word-ending analysis can also contribute for choosing the correct one.

2.2. Rule Based Segmentation

After a word is separated into stem and word-endings, the word-ending can further be segmented to singular suffixes by using singular suffix list, and by applying the phonetic rules which is necessary for recovering to standard surface forms. About 38,500 stems are collected which consists of almost all the common stems except from domain specific and rarely used stems. A relatively larger suffix list is obtained by applying these stems to a lexical corpus containing about 200,000 words. A list of compound and single suffixes and their corresponding maps are extracted. From the extracted suffix, 325 singular suffixes with a standard form of about 108 types are verified by manual checking, and about 5880 word-endings are automatically selected and corresponding singular components are split. Furthermore, new compound suffixes can be added automatically when the segmenter is trained on a new lexical corpus. As we can see that the stem and suffix boundary is the vital important for correct segmentation. Figure 2 shows the flow chart of the rule-based segmentation process.

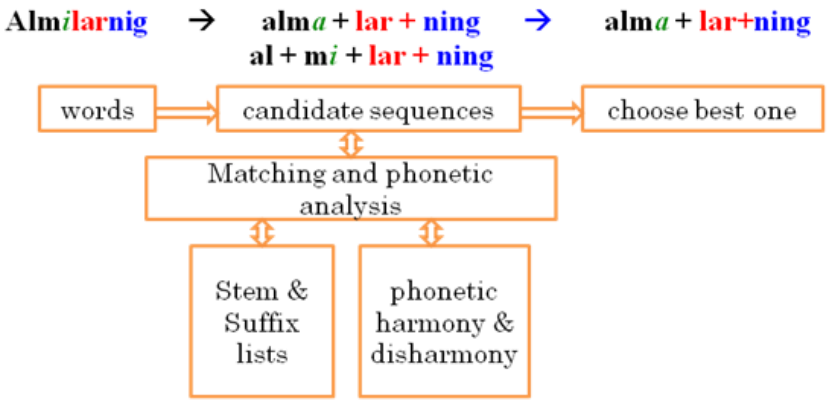


Figure 2. Flow Chart of Rule-Based Morpheme Segmentation

Matching algorithm can be applied iteratively for the segmentation of a candidate word which is chopped off gradually to match separately with the stem list and the word-ending list. When there are different segmentation results, the one with the longer stem is chosen to be the optimal. Choosing a longer stem decreases risk of incorrect segmentation, ambiguity, and confusion. For example word "almilarning (apples')" can be segmented to two segmentation results.

"almilarning = al (take)+milarning" before recovery,

"almilarning = alma (apple)+larning" after recovery.

We choose the longer stem as the preferred one because the word-ending of first one is incorrect or rare. For some OOV words, mostly imported from foreign

languages, which are not in the stem list, segmentation is carried out according to word-endings only, and incorrect segmentation may be produced, especially when the vowel weakening is happened. We selected 18,400 words from the text corpus for the evaluation, and split them to morphemes. After manually checking the segmentation result, we estimate the accuracy of segmentation is 88%.

However, this rule-based method has only a limited applicability, for the training and test sets are based on only a word list, lack of flexible segmentation ability, lack of context analysis, and almost ineffective for OOV words .

2.3 Morpheme Segmentation Based on a Statistical Model

A statistical learning model based Uyghur morpheme segmenter is developed based on a manually prepared training corpus of aligned word-morpheme sequences. 108 suffix types are defined and collected, according to their semantic and syntactic functions, which can be spread to 305 surface forms. In addition to the morpho-phonetic rules discussed in section 2.2, corresponding samples which are covered by the aligned training corpus are also learnt. All surface forms in the morpheme sequences are standardized before feed them to the statistical learning algorithm.

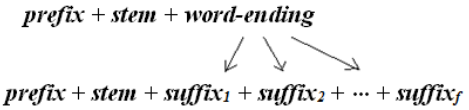
2.3.1 Corpus Preparation And Probabilistic Modeling

A text corpus of 10025 sentences and their manual segmentations are prepared and aligned to form a two-layer training corpus, statistical details are in Table 3. These sentences are collected from general topics, unrelated. Furthermore, more than 30K stems are prepared independently and used for the segmentation task. Only the singular suffix list with their standard forms is utilized for this method. Various surface forms of morphemes are learnt from the two-layer training corpus. Other surface forms that are not covered by the training corpus can also be predicted according to phonetic rules and tested according to the learning model.

Table 3. Statistics of Manually Prepared Morpheme Text Corpus

	tokens	vocabulary
word	139.0k	35.4k
morpheme	261.7k	11.8k
character	936.8k	34
sentence	10,025	

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 word ending (combined suffix or stem endings), and several possible stem word-ending pairs are obtained. Then, every word-ending is re-segmented into singular-suffixes, and sometimes each word-ending may have several different singular-suffix segmentations. All the possible segmentations of a word are extracted by matching based on various surface forms, and all probabilities are calculated in order to choose the best probable one. As we can see that the identification of stem and word-ending boundary is the most important part in segmentation.



There are ambiguity and confusion during the segmentation because of the reasons discussed in section 2.2, which are summarized in Table 4. First, phonetic harmony and weakening must be tested by recovering to standard surface forms. This is the main reason of causing different surface forms of a same morpheme. Second are the morphological changes like deletion and insertion. Third is the ambiguity of morpheme units, especially the stems.

Table 4. Examples of Problems in Morpheme Segmentation

segmentation example	problems
(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, bər(give)+ish	ambiguity

We use different solutions for the morphological and phonetic changes. For insertion, we add the inserted phoneme to the subsequent suffix as a new surface form. Thus stems are only suffered from deletion which can be learned from the training corpus. For the phonetic rules, we have to recover every of the 305 suffix forms into one of their original standard forms of 108 types, so that the suffixes are accurately segmented to their standard forms.

Generally, an intra-word bigram method based on the following probabilities is used, and the identification of stem-suffix boundary has a bigger impact on this segmentation

$$P(stem-suffix\ boundary)=\begin{cases} P(stem,firstSuffix)=\frac{C(stem,firstSuffix)}{C(stem)} \\ P'(stem)P(anySuffix|stem) \quad for\ smoothing \end{cases} \tag{3-1}$$

in which

$$P'(stem)=\frac{C(stem)}{stemToken+stemVocabulary} \tag{3-2}$$

$$P(anySuffix|stem)=\frac{C(stem,anySuffix)}{C(stem)} \tag{3-3}$$

where $C(t_i)$ is the frequency of t_i

2.3.2 Segmentation Results

We split the prepared corpus to the training corpus of 9025 sentences, and the test corpus of 1000 sentences. The results of coverage and segmentation accuracy are shown in Figure 3. 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 of automatic segmentation with morpheme of manual segmentation.

Figure 3 shows the graph of coverage and segmentation accuracy by various training corpus size. Morpheme units provide better coverage and smaller vocabulary size compared to word units. This segmentation tool can produce segmented morpheme sequence in both standard forms or in surface forms without degrading segmentation accuracy, so can be used for both speech and text processing applications.

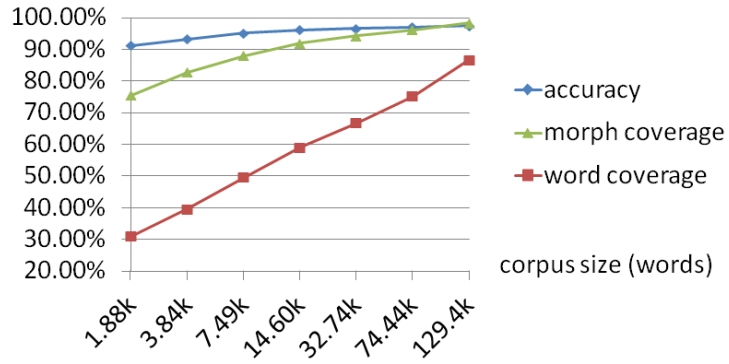


Figure 3. Morpheme Segmentation Accuracy and Coverage for Morphemes and Words

Observing the segmentation results, we can find the main causes of incorrect segmentation are the weakening problem, and ambiguity in the definition of the stem set. There are various segmentations of a same word in the learning corpus as in Table 4. However, this tool has only one segmentation result for a candidate word. The weakened stems (bar or bər) have an identical surface form when attached by certain suffixes. Both stems are frequent, but this tool can only produce the most probable one. Longer context analysis is necessary for solving the ambiguity problem.

3. Statistical Properties of Various Units

As an application of the morphological tool and a comprehensive part of Uyghur language compilation, we have tested the statistical properties of various lexical 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. They are from general topics such as novels, newspapers, and books (history, science...). This corpus is cleaned by removing all duplicated sentences, as it was a collection of different sources and may have many copies of same content. We segmented the texts in this corpus separately to morphemes and syllables, and built trigram language models based on three different units: word, morpheme, and syllable. All punctuation marks are removed in the following experiments to make the coverage and perplexity consistent in the language model (LM) evaluation.

We keep the surface forms of segmented morphemes same as in the words, thus the morpheme sequences can be recovered to words simply by re-connecting without any changes. This may cause some ambiguity in morphemes, but does not degrade segmentation accuracy. In this way, the statistical properties of coverage and perplexity are compared with n-gram language models.

Table 5. Example of Inserting Word Boundary Information for Various Units

Unit forms	People are unaware of the event.
Word sequence	Kishilər wəqədin bihəwər qaldi.
Morpheme sequence	Kishi _lər wəqə _din bi_ həwər _qaldi.
Syllable sequence	Ki shi lər _ wə qə di _ bi hə wər _ qal di.

In order to preserve the word boundary information, we adopt different methods for phonetic units and lexical units. For syllable and phoneme (character) units, a word boundary symbol is added between syllables or characters in the place of word boundary.

For morphemes, we label them as prefix, stem, and suffix, as shown in Table 5. In this way, the word sequences can easily be recovered from morpheme sequences by simply reconnecting them together.

As a test corpus, 11888 sentences are held out with the character size of 1460.8k, Table 6 shows statistics of the test corpus. From the statistics, we can see a word unit is segmented into about 1.88 morphemes and 2.73 syllables on average. The remaining 620K sentences are used as a training set. Trigram models are built on various units, respectively; Kneser-Ney smoothing is adopted. Unknown word model is <UNK> used, and words appeared only once are considered as unknown.

Table 6. Statistics of Test Corpus for N-Gram Evaluation

units	word	morph	syllable	statistical morpheme	character
tokens	217k	409k	593k	189k	1.4M
vocabulary	47k	15.3k	3.6k	43.4k	33

Figure 4-6 and Table 7 show the statistical results of vocabulary size, token, coverage, perplexity, and normalized perplexity. We can see that smaller units have better coverage and smaller vocabulary size. With larger n-gram dimensions smaller unit can have a better performance than words due to the low OOV rate.

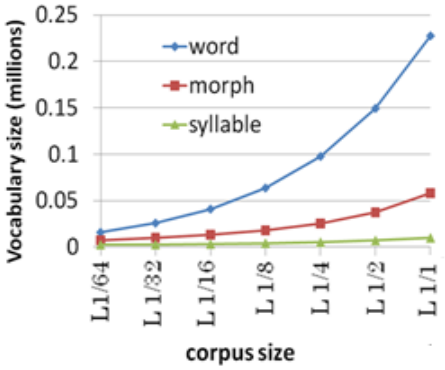


Figure 4. Vocabulary Size of Various Units

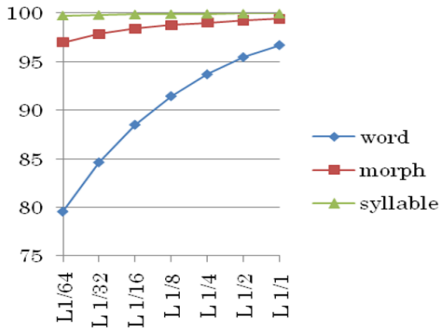


Figure 5. Unigram Coverage of Various Units

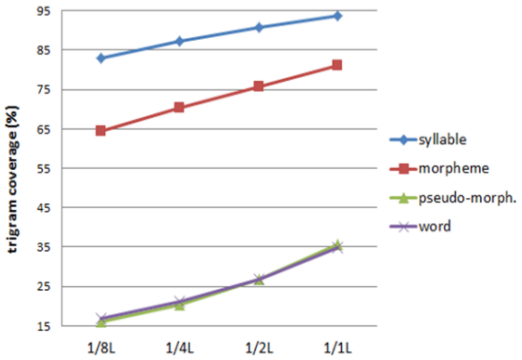


Figure 6. Trigram Coverage of Various Units

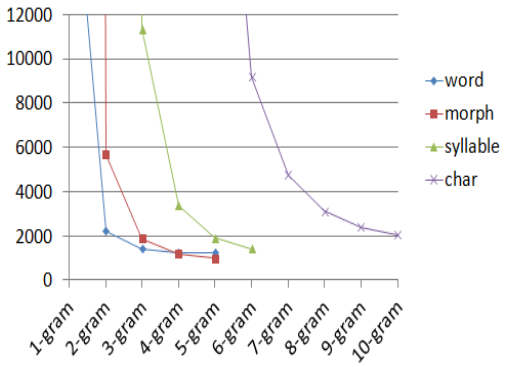


Figure 7. Comparisons of N-Gram Perplexities

Table 7. Comparison of Statistics and Trigram LMS Based on Various Linguistic Units

training corpus (sentences) L			77.5K	155K	310K	620K
subset of training corpus L			1/8	1/4	1/2	1/1
Vocabulary size	word	Freq>0	149,347	222,729	329,370	480,067
		Freq>1	63,501	97,461	149,054	227,101
	morph.	Freq>0	40,403	61,146	93,627	144,765
		Freq>1	17,823	25,145	37202	57768
	syllable	Freq>0	7049	9813	13,948	20,088
		Freq>1	4084	5180	6976	9846
#tokens		word	1,453,870	2,904,037	5,806,217	11,587,471
		morpheme	2,748,350	5,487,041	10,965,894	21,869,762
		syllable	3,987,209	7,959,562	15,906,966	31,744,522
character token			9,818,903	19,601,528	39,178,682	78,187,496
unigram coverage (%)		word	91.47	93.71	95.47	96.71
		morpheme	98.76	99.02	99.25	99.40
		syllable	99.87	99.90	99.93	99.95
bigram		word	53.07	58.64	64.56	71.10

coverage (%)	morpheme	89.60	92.07	94.08	95.77
	syllable	97.18	98.15	98.80	99.26
trigram coverage (%)	word	16.93	21.07	26.79	34.88
	morpheme	64.48	70.32	75.82	81.14
	syllable	82.84	87.24	90.80	93.69
perplexity	word	12856	6857	3689	1929
	morpheme	91.27	77.43	66.20	56.99
	syllable	26.13	24.12	22.52	21.228
normalized perplexity by word	word	12856	6857	3689	1929
	morpheme	5078.4	3706.8	2747.6	2060
	syllable	7699.7	6150.7	5076.3	4316.5

Table 8. Normalized Perplexity of N-Gram Models of Various Units

units	word	morpheme	syllable	character
vocabulary size	227.1k	57.7k	9.8k	33
1-gram	29,302	482,973	110,014,618	30,014,487,856
2-gram	3039	6294	168,482	140,025,078
3-gram	1929	2060	4316	4,498,647
4-gram	1754	1318	3349	217,874
5-gram	1700	1091	1901	29,051
6-gram			1425	9186
7-gram				4743
8-gram				3113
9-gram				2397
10-gram				2032

Then, we compare n-gram models with different n sizes, Table 8 and Figure7 show the results. 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. With larger n-gram dimensions, morpheme based model outperformed word based model, because of the low OOV rate. However, we can see that, with similar size of context, the various unit based n-gram models are converging to a similar perplexity value.

4. Conclusions

In this paper, we have discussed a general purpose morphological analyzer of Uyghur language, and proposed rule based and a statistical model based morphological unit segmentation approaches. During the designing and implementation of the supervised morpheme segmentation tool, we standardized and manually segmented the Uyghur morphemes, especially the suffixes, and summarized and classified morpho-phonetic rules and their implementations. 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.

By collecting large text and speech corpora, we have obtained reliable statistics for Uyghur language on various units as the comprehensive corpus compiling process of Uyghur language. Morpheme unit provides a small lexicon and better statistical

properties. It can provide a good foundation for downstream processing of natural language processing applications. Finally, this research provides a good example for the resource-scarce languages which also have agglutinative morphology.

Acknowledgements

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

References

[1] G. Adongbieke and M. Ablimit, "Research on Uighur Word Segmentation", Journal of Chinese information Processing, vol. 11, (**2004**).
[2] M. Ablimit, G. Neubig, M. Mimura, S. Mori, T. Kawahara and A. Hamdulla, "Uyghur *Morpheme-based* Language Models and *ASR*", In Proc. ICSP, Beijing, (**2010**).
[3] M. Ablimit, A. Hamdulla and T. Kawahara, "Morpheme Concatenation Approach in Language Modeling for Large-Vocabulary Uyghur Speech Recognition", In Proc. Oriental-COCOSDA Workshop, (**2011**).
[4] M.Y. Tachbelie, S. T. Abeta and L. Besacier, "Using different acoustic, lexical, and language modeling units for ASR of an under-resourced language- Amharic", Speech Communication, (**2013**).
[5] A. Lee, T. Kawahara, and K. Shikano, "Julius- an open source real-time large vocabulary recognition engine ", In Proc. Eurospeech, (**2001**), pp. 1691-1694.
[6] 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**).
[7] Graham Neubig, "Unsupervised Learning of Lexical Information for Language Processing Systems ", PhD thesis, Kyoto University, (**2012**).
[8] 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, (**2007**), pp. 1-29.
[9] 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**).
[10] M. Ablimit, T. Kawahara and A. Hamdulla, "Lexicon optimization based on discriminative learning for automatic speech recognition of agglutinative language", Speech Communication, (**2014**).
[11] M. Nuβbaum-Thom, A.E.D. Mousa, R. Schluter and H. Ney, "Compound Word Recombination for German LVCSR", In Proc. Interpeech, (**2011**).
[12] B. Xiang, K. Nguyen, L. Nguyen, R. Schwartz and J. Makhoul, "Morphological decomposition for Arabic broadcast news transcription", In IEEE-ICASSP, (**2006**).
[13] A. El-Desoky, C. Gollan, D. Rybach, R. Schluter and H. Ney, "Investigating the use of morphological decomposition and diacrit", In Proc. Interspeech, (**2009**).
[14] M. Jongtaveesataporn, I. Thienlikit, C. Wutiwiwatchai and S. Furui, "Lexical units for Thai LVCSR", Speech Communication, (**2009**), pp.379-389.
[15] T. Pellegrini and L. Lamel, "Using phonetic features in unsupervised word decompounding for ASR with application to a less-represented language", In Proc. Interspeech, (**2007**).
[16] E. Arisoy, M. Saraclar, B. Roark and 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**).

Authors

Mijit Ablimit, He 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, He 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.

Akbar Pattar, He received his B.E. degree in radio electronics from Xinjiang University, China, in 1983. He has been working as a teacher in School of Information Science and Engineering, Xinjiang University since 1983. His research interests include natural language processing and pattern recognition.

Askar Hamdulla, He 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 150 technical papers on speech synthesis, natural language processing and image processing. He is a senior member of CCF and an affiliate member of IEEE.