Partly Supervised Uighur Morpheme Segmentation

Mijit Ablimit
Graduate School of Informatics,
Kyoto University, Kyoto, Japan,
606-8501
mijit@ar.media.kyoto-u.ac.jp

Mihrigul Eli
Multilingual Information
processing Lab.
Xinjiang University, Urumqi,
China, 830046
mire1028@tom.com

Tatsuya Kawahara*
Graduate School of Informatics,
Kyoto University, Kyoto, Japan,
606-8501
kawahara@i.kyoto-u.ac.jp

Abstract· This paper introduces Uighur morpheme segmentation, which is a basic part of the comprehensive effort of the Uighur language corpus compilation, conducted at Xinjiang University in cooperation with Kyoto University. Uighur is an agglutinative language with word structures formed by productive affixation of derivational and inflectional suffixes to stems. Derivational suffixes change the meaning of the stems, while inflectional suffixes define grammatical functions, such as cases, of the stems. The surface realization of words is also constrained by phonetic rules such as phonetic harmony and vowel weakening, but the surface form of the stem is basically unchanged except for the last vowel. For example, the words "*adam+lar, adam+ni, adam+ga, adam+ning, adam+dak*" are formed by attaching different suffixes "*lar, ni, ga, ning, dak*" to the stem "*adam* (meaning person)". There are also complex suffixes or compound suffixes. They cause a huge number of combinations, thus the morpheme segmentation is the vital part of the Uighur language analysis.

We compiled lists of 38500 stems and 325 singular suffixes to cover most of general words. Then, a list of compound suffixes is collected in an unsupervised manner from our corpus of 200K words by matching with the basic list. With manual checking, 5880 compound suffixes were obtained. For automatic morpheme segmentation, we apply a forward and backward matching algorithm based on the list. One of the biggest problems is vowel weakening, that is, the last vowel of the stem "a" or "ä" is often replaced by another vowel "i" or "e". The phenomenon is observed for 12% of the words in our corpus. Thus, we have devised substitution rules, but these cause ambiguity in the morpheme segmentation. When more than one segmentation hypotheses are generated, the hypothesis with a longer stem is preferred; this is a safe heuristics.

Phonetic harmony is also a key factor that controls the stem-suffix connection and syllable concatenation. Thus, we have also introduced phonetic harmony rules which constrain the connection of the stems and suffixes in terms of the smooth articulation. For example, some voiced consonant at the end of a stem must be followed by a suffix starting with a voiced consonant. This constraint will effectively reduce the ambiguity.

The method was evaluated with 18400 words chosen from our corpus, and the accuracy of stem-suffix boundary detection is 96% and the accuracy of all stem/suffix segmentation is 92%. The result is encouraging since stems of some words, such as new words imported from English, are not included in the stem list. We are investigating an automated method based on a statistical model to cope with them.

Keywords: Uighur, morphology, phonetics

1 Introduction

Uighur (refers to the Uighur language) belongs to Turkish Language Family of Altaic Language system. It is an agglutinative language with word structures formed by productive affixation of derivational and inflectional suffixes to stems. Derivational suffixes are changing the meaning of the stems, while inflection suffixes are only changing grammatical functions of stems[1]. The surface realizations of morphological constructions are constrained and modified by a number of phonetic rules such as vowel weakening.

The work is funded by Natural Science Fund of China (No:60662002)

Uighur is written right-to-left in the Arabic alphabet with some modifications. There are 8 vowels and 24 consonants, total 32 letters in Uighur. Uighur morphology is an affixal system consisting mainly of suffixes and a few prefixes (6 in this research).

According to linguistic theory, morphemes are considered to be the smallest meaning-bearing elements of a language as well as the smallest units of syntax[18]. However, no adequate language independent definition of the *word* as a unit has been agreed on [2]. The task of morphological analyzer is to identify the lexeme, citation form, or inflection class of surface word forms in a language [2].

The utilization of morphemes as basic representational units in a statistical language model instead of words seems a promising course[8]. Many language processing tasks, including parsing, semantic analysis, information retrieval, and machine translation usually require a morphological analysis of the language beforehand.

The division of morphology and syntax in agglutinative languages is difficult. There is significant amount of interaction between morphology and syntax. Bound morphemes can indicate grammatical functions that are realized by words in languages like English [3].

2 Related works

There are several approaches for morpheme segmentation. Some of them are supervised and use some information and knowledge about the specific language such as morphological rules, stem list, suffix list, lexicon, etc[3][6][9]. Other approaches are unsupervised and use only a raw corpus to extract morphemes[2][4][8].

Some mathematical frameworks or modeling methodologies can be used for morphology learning and word segmentation: maximum likelihood (ML) modeling, probabilistic maximum a posteriori (MAP) models, finite state automata (FSA), etc. Despite the improvements in performance of the knowledge-free morpheme boundary detection, it is far below from what knowledge-rich system's performance.

Finite State Automata (FSA) can be used to describe the possible word forms of a language, for example, in the two-level morphology framework[19]. There exist algorithms that try to learn FSAs that compactly model the word forms observed in the training data, they require a segmented, and labeled corpus to begin with[7][9].

A stem centered segmentation method is proposed in this paper. The stem in Uighur remains fairly unchanged after suffixation, and makes this method relatively easier than suffix centered segmentation, considering needed manual works and complicated suffix structure.

Uighur has finite-state but rather complex morphological and phonetic rules. Morphemes (suffixes) added to a root or a stem can convert the word type (from a nominal to a verbal structure or vice-versa). The surface realizations of morphological constructions are constrained and modified by a number of phonetic rules such as vowel and consonant harmony and vowel weakening [1][13].

Some researches[16][17] have been done in Uighur morpheme segmentation, and claimed to achieve the accuracy of 85%[16]. But the specific morpheme segmentation methodologies are not discussed in details in these researches.

3 Inducing Uighur morphemes

Forward and backward algorithm is applied for the segmentation of a given word, and phonetic rules are examined for the words whose surface forms change after concatenation.

①phonetic rules: Extract the independent phonemes from the language, and analyze certain syllables, morphemes, and words according to the locations of phonemes.

②morphological rules: Extract morphemes from words. For example, the words "adam+lar, adam+ni, adam+ga, adam+ning, adam+dak" are formed by linking different suffixes "lar, ni, ga, ning, dak" to the stem "adam"(means: person, man).

3.1 Identifying stems and suffixes

Uighur morphology is complex and variable, influenced strongly by other languages, but never loses its integrity, preserving its intrinsic language rules. We focus on most general morphological rules which are common rules related to morpheme segmentation.

Surface forms of stems are relatively unchanged compared to suffixes when concatenated with other morphemes. In this research, at first, stems are collected manually from a dictionary [12]. To clarify, the root is the smallest independent meaningful unit; a stem is formed by linking derivational suffixes to a root. Derivational suffixes change roots (or stems) semantically while inflectional suffixes change grammatically. Therefore, stem list includes the roots as well. However, it is sometimes difficult to give a clear borderline for nouns that become verbs or vice versa. For example, root ”ish” is a noun, and a subject in a sentence. When the root is linked by different suffixes, syntactic or semantic changes happen. For example,

- ish+ lesh： “work”, become a verb， and a predicate in a sentence.
- ish+ci： “worker”, become a new stem.
- ish+tin：”from work”, can only be an adverb in a sentence.
- ish+ni： “the work”, can only be an object in a sentence.

To prevent over-segmentation and secure the semantic identity of a word, stem and suffix boundary is chosen as the primary target of segmentation. The segmented morphemes could further be segmented to roots and to singular suffixes automatically by using the stem list and singular suffix list. About 38,500 stems are collected as the basis of segmentation. The stem list consists of almost all the common stems except from the domain specific words and rarely used words.

A relatively complete suffix list was obtained in an unsupervised way by training these stems on a lexical corpus containing about 200,000 words. The training process was accomplished mostly by forward matching algorithm, because the stem list is the basis of the segmentation. A suffix list of compound and single suffixes are also extracted. Because of the final vowel weakening, the surface representations of the stems change when it is linked with suffixes.

From the extracted suffix, 325 singular suffixes are verified by manual checking, and about 5880 compound suffixes are automatically selected by segmenting to their singular counterparts. Furthermore new compound suffixes are added automatically when the segmenter is trained on a new lexical corpus.

Under the assumption that the stem list and the suffix list are the basis, a forward and backward algorithm is used to segment a candidate word. Sometimes when different segmentation results are come out, the result with the longer stem is chosen to be the output, as we choose the stem is the center of our segmentation. For instance, for the word "atamning", segmentation results can be "at+am+ning", "atam+ning". Only semantic or context analysis could find out that the second one is correct. Choosing longer stem decreases risk of incorrect segmentation.

4 Phonetic rules in Uighur

Phonetic rules in Uighur are based on the harmony of vowel, the harmony of consonant, and the final vowel change (weakening).

4.1 Final vowel weakening

When certain stems linked with some suffixes, the last vowel of the stem “a” or “ä” is replaced by two other vowels “i” or “e”; this phenomena is called “final vowel change (or weakening)” in Uighur language. Vowel weakening is a complex phenomenon. In a stem-suffix structure word, when the last syllable of a stem is accentuated, two neighboring accent impact on each other and cause weakening on the former one. Until now not a general formula is concluded to implement the weakening[1].

From a text corpus collected from newspapers and books, we extracted about 18,000 words, and vowel weakening is observed in about 12% words.

As we do not know when the weakening happens, it should be checked for every candidate word. Below are examples of final vowel change.

- maktipi=maktap+i , somebody’s school.
- adimi=adam+i, man from somewhere.

In this research, the method of solving the vowel weakening is to recover the weakened syllable. As we do not know which syllable is weakened, our method is to check one by one by recovering certain vowels.

After a candidate word is segmented to syllables, find letters “i” and “e” which may have been weakened, replace them separately with “a” and “ä”. Then recovered words can be segmented by forward and backward matching algorithm.

Several different segmentation results may be obtained. The stem can be over-segmented to a shorter stem and non-morphemes. For example, the word “almisi” (someone's apple) can be segmented to three different results: “almisi = alma + si”, “almisi = al + misi”, “almisi = almas + i”. In these, first and third are correct segmentations, only by semantic or context analysis can determine the correct segmentation, but choosing the longer stem is safer.

Because of the recovery process while dealing with vowel weakening, different segmentations may happen. For example word “almilarning” can be segmented to “al+milarning” before recovery, and segmented to “alma+larning” after recovery. In this situation, again we choose the longer stem as the preferred one. In the same time the suffixes analysis may also contribute to choose the correct one.

For the new words, mostly imported from other languages, which are not in the stem list, segmentation is carried out according to suffixes only, incorrect segmentation may be produced, especially when the vowel weakening is happened.

4.2 Syllable segmentation

A Uighur word consists of at least one syllable, and a syllable in Uighur contains only one vowel (except from some syllables imported from Chinese) and zero to four consonants. So the syllable number equals to the vowel number in a word. All syllables (except from words imported from other languages) in Uighur follow the rule: syllable=B+A+B+B (A is vowel, B is consonant) [13].

4.3 Phonetic harmony

There are two types of phonetic harmony in Uighur for the concatenation of vowels and consonants on the root-suffix interface.

①Rule of consonant harmony is the harmony of consonants according to the manner and point of articulation and the characteristics of the Uighur language.

②Rule of vowel harmony is the harmony of vowels according to the manner and point of articulation and the characteristics of the Uighur language[1].

Phonetic harmony is the basic controlling rule in the root-suffix linkage and syllable linkage. It happens at the interface of stem and suffix, and can be used to choose the correct form of a suffix. There are different forms of a same suffix in Uighur; only a certain form is used to link a particular stem according to phonetic harmony. There are four types of forms.

Type1: This kind of suffix, has only one form; it is not changed when linked to any stem, for example: “ni, ning”;

- adamning=adam+ning (correct)
- adamni=adam+ni (correct)

Type2: Consonants at the interface of stem and suffix must keep phonetic harmony according to surd or sonant. The consonants at the interface must be accordant with surd or sonant, for example, “din, tin”

- adamdin=adam+din (correct)
- adamtin=adam+tin (wrong)

Type3: Vowels at the interface of stem and suffix must keep phonetic harmony according to the articulation point. In this type, the vowels at the interface must be accordant with articulation point, for example, “lar, ler”

- adamlar=adam+lar (correct)
- adamler=adam+ler (wrong)

Type4: In this type, suffix form is chosen by both the type2 and type3, for example, “gha, qa, ge, ke”

- adamge=adam+ge (correct)
- adamgha=adam+gha (wrong)
- adamqa=adam+qa (wrong)
- adamke=adam+ke (wrong)

5 Experimental results

We implemented a morpheme segmenter based on the partly supervised method. In this approach a stem list is the basis of segmentation. Our corpus contains 38,500 stems, 325 singular suffixes, and about 5880 compound suffixes. 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 the segmentation. The accuracy of the detection of stem-suffix boundary is above 96%, and the accuracy of further split to singular suffixes is 92%.

Different evaluation measures can be used, for example, precision rate, recall rate, F-measure. But, in this research, stems included in the stem list have more advantage than the stems not yet included. And, even for the non-included words a segmentation result is obtained, in which at least a stem or a suffix is correct.

6 Conclusion

Suffixes in Uighur are complex, especially when a stem is linked with many suffixes. For example: “ishcilarningki = ish+ci+lar+ning+ki”. The linkage of suffixes between them and their order are complex, and yet to be studied.

The new words not included in the stem list must be added manually or automatically by some unsupervised statistical analysis[4][5]. When this algorithm incorporated with some specific applications, like spell checker or search engine, revisions may be needed according to specific applications.

References

[1] Hamit Tomur. *Modern Uighur grammar (Lexical study) [M]*. Beijing, National Publishing House of China, 1987.

[2] Mohsen Arabsorkhi, Mehrnoush Shamsfard. *Unsupervised discovery of Persian Morphemes*. 11th International CSI Computer Conference, Tehran, Feb. 2006

[3] Burcu Karagol-Ayan. *Morphosyntactic generation of Turkish surface forms*. Proceedings of the ESSLLI student Session 1999.

[4] Stefan Bordag. *Unsupervised and knowledge-free morpheme segmentation and analysis*. morphochallenge, 2007

[5] Mathias Creutz, Krista Lagus. *Unsupervised models for morpheme segmentation and morphology learning*. ACM Transactions on speech and language processing, Vol.4, January 2007.

[6] Kemal Oflazer. *Two-level Description of Turkish Morphology*. Literary and Linguistic Computing, 1994

[7] Richard Sproat. *Book Reviews PC-KIMMO: A Two-Level Processor for Morphological Analysis*. AT& T Bell Laboratories, 1990

[8] Mathias Creutz and Krista Lagus. *Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor1.0*. In Publications in Computer and Information Science, ReportA81, Helsinki, Finland, March2005. Helsinki University of Technology.

[9] Hans J. *Nelson. A two-level engine for tagalong morphology and a structured XML output for PC-KIMMO*. Mater’s Dissertation. Department of Linguistics and English Language, Brigham Young University, August 2004.

[10] Gulila Adongbieke. *The Research of Proofreading for the Uighur Character.* The 2001 IEEE International Conference on System, Man and Cybernetics (SMC2001), 2001.10.7-10.10, Tiscon, Arizona, U.S.A, P874-876.

[11] Mijit Ablimit, *Research on Uighur corrector system in multilingual environment[C]*. System engineering theory and practice”. 2003,23/5: P117-124.

[12] *Uighur spelling and pronunciations dictionary [M]*. Urumqi, Xinjiang People's Publishing House, 1997.

[13] Mijit Ablimit, Askar Hamdulla, Kurban Ubul. *Phonetic harmony in Uighur Language and its Implementation*. CSTA 2005 Association, August 2005.

[14] Mayire Yibulayin, Mijit Ablimit, Askar Hamdulla. *Generation and Correcting of Spelling Errors in Uighur Based on Minimum Edit Distance*. Journal of Chinese information Processing, May 2008.

[15] Turdi Tohti, Winira Mosajan, Mijit Ablimit. *Uighur Search Engine Web Server and its Implementation*. The 11th National Minority Language Information Processing Conference Proceeding. January 2007.

[16] Gulila Altenbek. *Automatic Morphological Tagging of Contemporary Uighur Corpus*. Proceedings of the 2006 IEEE International Conference on Information Reuse and Integration, IRI – 2006,Hawaii, USA. IEEE Systems, Man, and Cybernetics Society 2006.

[17] Yusup Abaidula，Rezwangul, Abdiryim Sali. *The Research and Development of Computer Aided Contemporary Uighur Language Tagging System*.

Journal of Chinese Language and Computing. 2005.

[18] Matthews, P. H. *Morphology.* 2nd edition. Cambridge, England: Cambridge University. 1991

[19] Koskenniemi, K. *Two-level morphology: A general computational model for word-form recognition and production.* University of Helsinki, Department of General Linguistics, Publications, No. 11. 1983.