Using a Model of Scripts for Morphological Segmentation Given an Unannotated Corpus

Anil Kumar Singh and Harshit Surana
Language Technology Research Centre
International Institute of Technology, Hyderabad, India
anil@research.iiit.net, surana.h@gmail.com

1 Abstract

Morphological analysis is an important step in natural languages processing which can not only increase the performance for many problems like shallow parsing, but can be compulsory for deeper syntactic or semantic analysis. In our view, morphological analysis can be divided into two stages, namely morphological segmentation and assigning morphological features. In this paper we present an *akshar* based approach for the first stage, i.e., morphological segmentation. The reason we are using an *akshar* based approach is that most of the major South Asian languages use scripts derived from Brahmi and are *aaksharik* in nature, i.e., *akshar* is the orthographic unit in these scripts. The only resource that we use is an unannotated corpus. We first build a list of word types from this corpus and at the same time compute their frequencies. Then this list of word types is compiled into forward and backward *tries* of *akshars*. Note that the nodes of the tries are *akshars*, not letters. Some features are also filled in for the nodes, e.g., nodes which mark the end of a word type will have a word-end feature turned on. Since an *akshar* may include part of an affix, we also use addition (moving up in the trie) and deletion (movind down the trie) operations to find out where the affix starts or ends. By calculating the frequencies of common substrings, we try to predict the morpheme boundary. For this purpose, three kinds of substrings are considered differently: word beginning, word ending and word middle. Since our approach is based on a common model of Brahmi origin scripts, it can be easily applied to all the languages which use these scripts. The advantage of our method is that we benefit from the similarities among South Asian languages (and their scripts) and use abstract linguistic knowledge without actually relying on manually created resources. At the same time we also use statistical information obtained from the unannotated corpus.

Keywords: Model of scripts, morphological analysis, morphological segmentation, Brahmi origin scripts, *akshar*, South Asian languages