Design and Implementation of Algorithms for Morphology Learning and its Applications

By

Shilpa Neenad Desai

Abstract

A Morphological Analyzer is a tool that identifies morphemes in an input word and outputs a lemma along with analysis of the morphemes. It is a basic building block used in the creation of complex Natural Language Processing (NLP) tools for a language, such as Spell Checkers, Text to Speech Systems, Cross-Lingual Information Retrieval, Rule Based Machine Translation etc. Developing a Morphological Analyzer comprises of two distinct tasks, identification of Morphological Paradigms and generation of a Morphological Lexicon. Automation of these tasks is generally referred to as Morphology Learning. The objective of this research work was to develop NLP tools and resources for Konkani. Konkani did not have a Morphological Analyzer and also lacked resources such as Suffix List or Morphological Paradigms, which are required to build a Morphological Analyzer. The lack of these resources formed the motivation to design and implement algorithms for extracting Suffix List and Morphological Paradigms; generating Morphological Lexicon and creating Finite State Transducer for Konkani.

To automate the task of extracting Suffix List and Morphological Paradigms, we designed an algorithm using the proposed concept of Suffix Association Matrix (SAM). This algorithm, which we have referred to as SAM based Morphology Learning algorithm, is designed to prune invalid segmentations caused due to short stem length words by modifying the existing p-similar technique. The Signature Repository generated using SAM algorithm was used to create the initial seed set for the Suffix List and the initial Morphological Paradigm List. These Lists generated using Unsupervised Morphology Learning approach were enhanced by referring to grammar books,
Generating a Morphological Lexicon, which maps lemmas to appropriate Morphological Paradigms, would be time consuming and expensive if done manually. Automating this task was a plausible solution, but not a straightforward task, since there are no linguistic rules to map lemmas to their corresponding paradigms. Hence Automatic Paradigm Selection (AutoParSe), an algorithm for automatically selecting paradigms was designed. AutoParSe used the proposed concept of paradigm differentiating measure \((pdm)\) to select multiple paradigms for a lemma and made use of logistic regression to distinguish between paradigms that differ only at the lexical level. The paradigms created were used to extract morphological features and these features were used in developing a staged approach for Grammatical Gender Identification of nouns in Konkani.

The Morphological Lexicon developed using AutoParSe, was further enhanced using the output of the Lemma Extraction algorithm wherein Paradigm Relevance Constraints were designed for each paradigm to extract additional lemmas from the Corpus. The Morphological Paradigm List and Morphological Lexicon were used to design and implement the FST based Morphological Analyzer for Konkani using foma, a finite state compiler. Using the FST based Konkani Morphological Analyzer a Web Based Spell Checker has been developed. The Konkani Morphological Analyzer is currently made available through a web based interface.

The algorithms designed and implemented as part of this research can be used to develop Morphological Analyzer for other languages as well by providing the required morphological inputs of the respective languages. The output of this doctoral research work will prove very useful for future work in NLP, especially for Konkani Language. An immediate application would be to enhance resources like the WordNet and the Corpus developed for Konkani.