This thesis focuses on designing data-driven Natural Language Processing solutions for a low-resource language like Sanskrit. Our design decisions are motivated by taking the language's typology, usage and resource availability into consideration. The culture bearing language of India has about 30 million extant manuscripts that are potent for digitisation. Languages across the world exhibit wide typological variations, while the design decisions for modelling NLP solutions often are based on a few resource rich languages. In fact, many of these assumptions need not even be valid for a language like Sanskrit.

First, we show that we can successfully combine linguistic information from Sanskrit grammar, lexical networks and distributional information into data driven models for word segmentation and tasks pertaining to word-formation. We identify tasks, such as compound type identification and identification of derivational nouns, which are often overlooked in other languages, but are essential for processing Sanskrit texts due to the typological characteristics of the language. We also build a word segmentation model where linguistic information is incorporated, both in constructing the search space as well as in feature engineering. In the second part of the thesis, we investigate the applicability of purely data-driven approaches for sequence level tasks in Sanskrit, especially under low resource settings. Seq2seq model based solutions are proposed for these tasks, where we compensate for the lack of training data by synthetic generation of training data and augmenting training with auxiliary resources. Here, we formulate the task of converting the word order in a verse to its corresponding prose order as a linearisation task. This enables the use of prose-only data as auxiliary training data, thereby avoiding the need for parallel data in verse and prose order.

Finally, we propose a generic graph-based parsing framework using energy-based models for multiple structured prediction tasks in Sanskrit. We experiment with the tasks of word-segmentation, morphological parsing, dependency parsing, syntactic linearisation and prosodic linearisation. For a free word order language like Sanskrit, we observe that the graph based non-sequential processing of input outperforms sequential processing approaches even for low level tasks such as word segmentation and morphological parsing. We automate the learning of the feature function which, along with the search space, encode relevant linguistic information for the tasks we consider. This effectively results in use of, as low as 10 % of the task-specific training data as compared to data requirements of the current neural state of the art models for various tasks.

The major contributions of the thesis include solutions to a range of challenges in analysing Sanskrit. We observe that integration of linguistic knowledge, both in designing search space and feature function, substantially reduces the task-specific training data requirements, a desirable aspect for low-resource languages. We also introduce two novel tasks of verse to prose order conversion and prosodic linearisation, which can aid digitisation and processing of Sanskrit texts. In all the tasks we discuss, we achieve state of the art results and in some of the tasks, ours is the only data driven solution.

Keywords: Sanskrit Computational Linguistics, Morphology, Syntax, Low-resource Language, Word Segmentation, Morphological Parsing, Dependency Parsing, Syntactic Linearisation, poetry linearisation, Word Formation, Compound Type Identification, Derivational Nouns, Post-OCR Text Correction. Structured Prediction.

Shorter Version

This thesis focuses on designing data-driven Natural Language Processing solutions for a low-resource language like Sanskrit. Our design decisions are motivated by taking the language's typology, usage and resource availability into consideration. Typically NLP solutions, which are often designed with simplifying assumptions based on characteristics of a few resource rich languages, may not be directly usable in the case of Sanskrit. First, we show that we can successfully combine linguistic information from Sanskrit grammar, lexical networks and distributional information into data driven models for word segmentation and tasks pertaining to word-formation. We identify tasks, such as compound type identification and identification of derivational nouns, which are often overlooked in other languages, but are essential for processing Sanskrit texts due to its typological characteristics. In the second part of the thesis, we investigate the applicability of building purely statistical approaches for two sequence level tasks in Sanskrit, especially under low resource settings. Seq2seq model based solutions are proposed for these tasks, where we compensate for the lack of training data by synthetic generation of training data for Post-OCR text correction and augmenting training data with auxiliary resources for converting the word order in a verse to its corresponding syntactic order. Finally, we propose a generic graph-based parsing framework using energy-based models for multiple structured Prediction tasks in Sanskrit. We experiment with the tasks of word-segmentation, morphological parsing, dependency parsing, syntactic linearisation and prosodic linearisation. For a free word order language like Sanskrit, we observe that a non-sequential graph based processing of input outperforms sequential processing approaches even for low level tasks such as word segmentation and morphological parsing. We automate the learning of the feature function which, along with the search space, encode relevant linguistic information for the tasks we consider. This effectively results in use of, as low as 10 % of the task-specific training data as compared to data requirements of the current neural state of the art models for various tasks.

The major contributions of the thesis include solutions to a range of challenges in analysing Sanskrit. We observe that integration of linguistic knowledge, both in designing search space and feature function, substantially reduces the task-specific training data requirements, a desirable aspect for low-resource languages. We also introduce two novel tasks of verse to prose order conversion and prosodic linearisation, which can aid digitisation and processing of Sanskrit texts. In all the tasks we discuss, we either achieve state of the art results or ours is the only data driven solution for the task.