Preface

This thesis investigates the machine learning based approach for analyzing the Hindi compound noun semantics. Noun compounds are interesting constructions from the perspective of theoretical linguistics and computational linguistics for the complex semantic relations between their constituents. The meaning of Compound Noun is composed of the meanings of the individual constituents and the way they are semantically related. Noun compound interpretation is the task of detecting this underlying semantic relation. For instance, a kitchen knife is a knife which is used in the kitchen (used in or purpose relation) whereas a steel knife is a knife made of steel (made of or constituent relation). This work explored different machine learning techniques as well as a knowledge base method which can predict the semantic relations between the constituents of compound nouns.

We developed a semantic relation set to annotate the semantic relations between the constituents of compound nouns,describing in detail the motivation for the relation set and the development process. Using the semantic relation set an annotated dataset is created to use as an experiment dataset with different machine learning algorithms. The inter-annotator agreements result indicate the reliability of the semantic relation and dataset for Hindi Compound Nouns analysis.

We treated compound noun interpretation problems as a classification problem for ML tasks and experimented with SVM and Random forest. Hindi WordNet is used as a linguistic resource and has got a significant output. We also experimented with Embeddings, since embedding captures more semantic features in language models and gives better results.

In this work, we also attempted to create a knowledge base in the domain of Ayurveda using the Generative Lexicon framework of lexical knowledge representation. Ayurveda text has a very limited corpus. Therefore, using a probabilistic model does not work. We developed a lexical knowledge representation database of some selected frequent nouns of Ayurveda corpus using the Generative Lexicon framework. The GL model is able to represent the natural polysemy of the word in the lexicon which is used for disambiguation tasks. Language models with rich knowledge encoded resources can be beneficial for developing linguistically precise probabilistic models.