PREFACE

From the initial days of the World Wide Web (the 1990s) till date, the Web primarily comprises English data. The data available on the Web is predominantly unstructured text. This unstructured data requires efficient linguistic tools and pre-processing steps to improve the effectiveness of the text analysis tasks. Several linguistic tools are proposed, such as tokenizers, stopwords, stemmers, lemmatizers, POS taggers, named entity recognizers and morphological analyzers in English text-processing tasks. These tools substantially improve effectiveness in the text-analysis domain. Some of these linguistic tools are also used as the components of machine translation, information retrieval and other NLP systems. Although different linguistic tools and pre-processing strategies have been proposed and evaluated in English, the domain is less explored in low-resource Indian languages. The primary reasons have been less availability of linguistic resources on the Web and technically skilled human resources.

There has been a rapid proliferation of information technology in the recent past, and ubiquitous computing has spurred tremendous growth of text data being captured in digital form on the Web. This text data comprises multiple low-resource languages other than English. Some of the low-resource languages are morphologically rich and have different kinds of lexical, morphological, syntactic, and semantic variations. In some low-resource Indian languages, few linguistic tools such as lemmatizers, POS taggers, named entity recognizers and morphological analyzers are available. The unavailability of linguistic resources and tools reduces the effectiveness of computational tasks such as machine translation, speech recognition, sentiment analysis, topic modelling, text classification and information retrieval (IR). In recent years, different researchers have proposed linguistic tools for various Indian languages. Although few linguistic tools are proposed in the text analysis domain, they are not evaluated from NLP and IR perspectives. This thesis investigates three pre-processing strategies: stopword removal, stemming and de-

compounding techniques in different Indian languages IR.

At first, we studied the impact of different stopword lists (non-corpus-based and corpus-based) in different Indian languages IR. Experiments on various Indian languages such as Bengali, Marathi, Gujarati, Hindi, Sanskrit and English show that stopword removal generally improves mean average precision (MAP) score compared to the case when it is not done. We also notice that the corpus-based stopword list outperforms the non-corpus-based stopword list. Different lengths of the stopword list are explored and evaluated for each language, which leads to suggesting its optimal length. We observe that a smaller length of a corpus-based stopword list outperforms the larger length of a non-corpus-based stopword list in the Indian language IR. We also investigate the effect of stopwords on retrieval effectiveness over document length. The impact of stopwords is relatively low in short documents compared to their long counterparts across the Indian languages.

Sanskrit is an ancient and sacred Indian language but less studied computationally. We thus build a Sanskrit text collection and explore different indexing, stemming and searching strategies. Stemming is an essential pre-processing step in the text analysis domains such as text mining, text summarization and IR. In this study, we propose two stemmers: a 'light' and an 'aggressive' and evaluate their effectiveness in the text processing task. Experiments on Sanskrit text collection show that the stemming method improves the effectiveness of an IR system. Among the different stemming techniques, the 'aggressive' strategy performs best in the text analysis domain.

As the final contribution, we investigate the effect of decompounding models in the Indian language IR. In this study, we propose different corpus-based, hybrid machine learning-based and deep learning-based decompounding models and evaluate their effectiveness in the IR domain. Experiments on different Indian languages (Marathi, Hindi and Sanskrit) show that the different decompounding models improve the effectiveness of an IR system. Among the different decompounding models, the attention-based deep learning models outperform the corpus-based and hybrid machine learning-based models in Indian language IR.

We perform extensive simulations on different pre-processing strategies on standard IR settings and measure their retrieval effectiveness in different metrics: precision, recall, R-prec, average precision, and mean average precision. We show the simulation results

on standard benchmark data sets. We also compare the results with different baseline approaches and algorithms and present them in various tabular and graphical forms for comparison. The results demonstrate that the pre-processing strategies have substantially improved the effectiveness of IR systems.