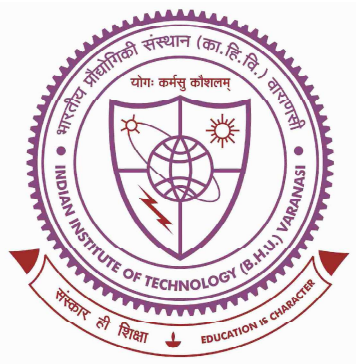
Pre-processing in Indian language IR



Thesis submitted in partial fulfilment
for the award of degree

Doctor of Philosophy

by

Siba Sankar Sahu

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
Indian Institute of Technology
(Banaras Hindu University)
Varanasi

Roll No: 17071012

Year of Submission: 2024

Chapter 9

Conclusions and Future Work

9.1 Summary and Contribution

Pre-processing is an important step in the text analysis domain. Although the effect of pre-processing strategies is extensively studied in the European languages, it is less explored in the Indian language IR. This thesis studies different pre-processing steps for information retrieval tasks in a few Indian languages. We considered three main pre-processing steps: stopword removal, stemming and decompounding. We evaluated the effectiveness of different pre-processing steps using standard measures.

Chapter 1 began with the guiding motivation, introduced the issues of different preprocessing steps, and highlighted the contributions of the thesis. Chapter 2 presented a thorough literature survey on pre-processing strategies in the text analysis domain. Next, Chapter 3 presented the simulation environment, datasets and evaluation metrics used for evaluating the pre-processing tasks in the Indian language IR.

In Chapter 4, we evaluated the effect of stopword removal in retrieval using different non-corpus-based stopword lists for the Indian languages. The stopword lists are downloaded from GitHub sources, and experiments are conducted from the IR perspective. Experiments on the Indian languages show that stopword removal improves MAP, R-prec, and precision@10 scores compared to their without stopword removal counterparts. Further, we investigated the relationship between stopwords and average document length. The effect of stopwords is relatively low in short documents compared to long ones.

In Chapter 5, we evaluate the effect of different non-corpus-based and corpus-based stopword removal in the Indian language IR. The corpus-based stopword lists are generated by applying different statistical approaches. Experiments on the Indian languages show that both

non-corpus-based and corpus-based stopword removal improves MAP scores with respect to the baseline (without stopword removal). A smaller corpus-based stopword list outperforms a larger non-corpus-based stopword list from a retrieval perspective. We also suggested an optimal length stopword list to be used for different Indian languages.

In Chapter 6, we built a text collection for Sanskrit and presented the morphological difficulties in Sanskrit text processing. We also proposed two rule-based stemmers, where one (called 'light') strips the inflectional suffixes, and another (called 'aggressive') strips both inflectional and derivational suffixes. Among the stemming strategies considered, the aggressive stemmer performs the best from both NLP and IR perspectives. We also demonstrated that the language-independent (trunc-n) indexing approach provides a similar MAP score to the best-stemming approach experimented with.

In Chapter 7, we proposed different corpus-based, hybrid machine learning-based and deep learning-based decompounding models and evaluated their effectiveness in the IR domain. Experiments on Indian languages showed that the different decompounding models improve MAP scores and the number of relevant retrieved documents compared to their no-decompounding counterparts. The corpus-based decompounding models perform poorly compared to the other decompounding models that were experimented with. The attention-based decompounding models provide the best MAP score in the Indian language IR.

9.2 Possible Research Directions

As discussed earlier, the thesis studies different pre-processing steps from the perspective of Indian language IR. There can be several directions of follow-up work that can be taken up. Here, we propose some exciting and promising problems for future research in pre-processing strategies in the text analysis domain.

9.2.1 Building text collection

- Despite the fast proliferation of low-resource languages on the Web, no standard text collections are available for low-resource Indian languages such as Tamil, Telugu, Malayalam and Kannada. In future, steps can be taken to build text collections for such languages.
- The languages mentioned above also suffer from good linguistic resources like POS tagger, morphological analyzer, stemmer, lemmatizer, etc. Initiatives are also needed to build these tools.

9.2.2 Stopwords

- We intend to experiment with the effect of stopwords in other Asian language families, such as the Dravidian language family. The Dravidian language family comprises Tamil, Telugu, Malayalam and Kannada. Researchers can explore the effect of stopwords in these languages.
- Machine learning and deep learning-based stopword detection techniques can be explored in Asian and European languages.
- One of the future directions can be exploring zero-shot and few-shot techniques to detect stopwords in Asian and European languages.

9.2.3 Stemming

- The effect of stemming strategies can be explored in a larger Sanskrit text collection.
- A hybrid and unsupervised stemming technique can be explored in low-resource Indian languages.
- A machine learning and deep learning-based stemming technique can be explored to generate root words in Asian and European languages.

9.2.4 Decompounding

- In future, the decompounding model can be used in NLP-related tasks in low-resource Indian and European languages.
- An advanced neural network model can be explored as decompounding models for low-resource languages.

9.2.5 Other Text Processing Applications

- The effect of pre-processing strategies can be explored in other computational tasks like machine translation, speech synthesis, text classification and text summarization.
- In future, the researchers can explore the effect of pre-processing strategies in neural IR.