Chapter 6

Conclusion and Future Enhancement

6.1 Summary

Pre-processing of any Text Mining and Natural Language Processing application requires a very important task to normalize the words before going further with actual implementation. Through literature survey it was observed that the popular stemmers normally generate stems which may not be actual dictionary words. This could lead to inaccurate or incomplete output in applications related to TM and NLP.

The detailed study of existing lemmatizers and their analysis showed that correct lemmas are not generated for the allied morphed words and nominalized words. This could also result in loss of correct and precise output. This was the research gap which was handled in the work described in this thesis. In retrospect, the problem statement and research objectives as mentioned in Chapter 1 and how they have been attained is described briefly below.

The problem statement was as follows:

'The aim of this study is to design, develop and implement a Lemmatizer for English morphed words handling nominalization and giving a better performance and output as compared to the existing popular lemmatizers.'

As per the problem statement, detailed literature study was conducted and two lemmatizers were developed. Along with this, each objective of the research study was also achieved. The objectives thought about at the beginning of this work and how they were fulfilled is explained in the next section.

6.2 Outcome of the Research work

The objectives of this research work were to develop a simple, robust and enhanced model for lemmatization. The model is supposed to overcome the shortcomings of the existing stemmers / lemmatizers and generate an accurate, precise and exact output in terms of lemmas for the input text. Each objective along with the attainment is described below.

Objective: 1

'*To study and to implement existing stemming and lemmatization approaches and to compare the limitations and advantages of each'*.

Attainment: 1

To achieve the above-mentioned objective, research papers related to stemming and lemmatizations were deliberated upon.

The algorithms of popular stemmers like Lovins, Porter, Paice, KStemmer and Bi-gram Statistical stemmers were studied and implemented. The analysis of the correctness and strength of each stemmer were observed and detailed comparative summary was done.

The prevalent lemmatizers like Stanford, Lemmagen, Wordnet and spaCy etc. were executed and their output were also summarized in detail. The analysis revealed the limitations of these lemmatizers.

This has been covered in Chapter 2 and Chapter 3 of the thesis. Two survey papers related to the work done were published in SCOPUS indexed journals. The details of the papers are given in the publication section.

Objective: 2

'To incorporate grammatical word formation rules for constructing English derivative words and to apply statistical distance measures to minimize the erroneous result and limitations of the existing stemmers and lemmatizers'.

Attainment: 2

To understand the morphological structure of English surface / inflected / derivative / nominalized words, different research papers and books related to morphological analysis and NLP were studied. Based on the detailed study and interaction with computational linguistic experts, exhaustive list of suffix-rules and corresponding recoding rules were constructed. This detailed exploration formed the base of the two models of lemmatization which were developed later. This has been discussed partly in Chapter 1 and Chapter 4.

Objective: 3

‘*To design a lemmatizer which generates correct lemmas for different morphed, derived and nominalized words and comparing them with the standard and popular lemmatizers. The lemmatizer should work on the text as a whole combining related morphed /derived words making the processing and output simpler and acceptable’*.

Attainment: 3

Two models related to lemmatization-named as LemmaChase and LemmaQuest were designed and implemented to fulfill the research gap and attain the problem statement as well as the third research objective. These models were compared with existing lemmatizers to show their better efficiency in generating the correct lemmas for English morphed words and also handling nominalization. Papers related to this work were also published in SCOPUS indexed and UGC CARE journals as mentioned in the publication section.

6.3 Research Output

Two models of lemmatization for English morphed words handling nominalization are presented and implemented that offer a generic and logical solution to meet the research objectives.

The first proposed model-LemmaChase provides a morphological analyzer in which word formation rules have been incorporated to generate a dictionary word from a derived input word. It also handles inflected input words properly just like other lemmatizers. The lists of lemmas which are extracted for the input are more accurate and error-free as compared to lemmas generated by the existing available lemmatizers.

The second proposed model-LemmaQuest is very capable and efficient when the input text has a large number of allied morphed words which coexist in that text. This lemmatizer works on the whole input text instead of executing on each derived or morphed word separately. This model generates groups of allied words dynamically to minimize the lemma extraction processing task. It generates a single lemma for each group that gets created. The model segments words to generate stem-token which is further processed to generate the correct lemma.

After the execution of LemmaQuest, it is observed that the output generated by it is far more accurate, and the error rate is far less as compared to the output generated by other existing lemmatizers as well as the previous LemmaChase model. The detailed comparison between LemmaChase and LemmaQuest is discussed in Chapter 5.

Any TM or NLP application which implements LemmaChase or LemmaQuest (depending on the type of input) in the pre-processing step, would give a better result as compared to that when using a currently prevalent stemmer or lemmatizer. This has been shown in detail in Chapter 4 and Chapter 5.

6.4 Future Enhancement

Designing and developing an efficient lemmatizer for English morphed words handling nominalization gave an insight into how this work could be carried further and how interesting it can be for other researchers working in this area. This research could be considered as a stepping stone to extend it as part of post-doc work also.

The suggestions for future enhancement are as follows:

a. These two models can be extended to become generic for the processing of words for regional languages or foreign languages too.
b. A lemmatizer model handling the prefixes of words and relating the lemma with the context of it being a possible antonym of the prefix morphed word could be developed.
c. The lemmatizers designed and developed work on the concept of words which are analyzed syntactically independent on the word’s context in the sentence. It would be a challenging task if the models are extended to analyze the words at semantic levels for better understanding of the meaning of the input text. This could be a very fascinating and thought-provoking research work.

This concludes the detailed write-up of dissertation related to the research work done. The publication details related to this work, the references with related content follows in subsequent sections.