ABSTRACT

Morphological Analysis in linguistics is the study of the principles and processes by which words are constructed and analyzed in any language. It specifically refers to extracting the significant and meaningful parts from words. Stemming and lemmatization are major morphological analyzing operations that are used as pre-requisite tasks for most of the Natural Language Processing (NLP), Text Mining (TM), Information Extraction (IE), Information Retrieval (IR) etc. related applications.

Stemming is a primitive way to obtain stem-token by chopping off the ends of any morphed word which are nothing but inflectional or derivational suffixes. After execution of any existing popular stemming process, it is observed that stems of morphed words are not necessarily dictionary words. Alternatively, lemmatization obtains the root or lemma of an input morphed surface-word which is an actual dictionary word. Therefore the stems generated by stemming cannot be used effectively as an informative entity in NLP or TM, since they do not make sense as compared to lemmas generated through lemmatization.

Meaningful words are absolutely important for applications like text simplification, key-term identification, question-answering systems, text summarization, topic identification, etc. in morphologically rich languages. This is especially in cases where one root word might have many morphological variants due to agglutination or inflection. In an IE or IR system, therefore lemmatization can provide support to improve overall retrieval recall since a query will be able to retrieve more relevant documents when variants in both query and documents are morphologically normalized. The research gap after extensive literature study was found in the generation of correct lemmas and hence the need of an efficient lemmatizer was realized.

This research work proposes two lemmatization models which are designed based on both, stemming and lemmatization techniques for obtaining the correct lemma from allied morphed words present in any input text. These models significantly minimize the limitations of the currently available popular stemmers like Porter, Lovins, Paice, YASS etc. and lemmatizers like the Stanford-LemmaProcessor, spaCy Lemmatizer, LemmaGen, WordNet Lemmatizer, Morph-Adorner etc. The existing lemmatizers generate correct lemmas for all inflected English morphed words, but not for any type of derived words; especially for nominalized derived words where POS of derived word and root word are different. The concept of nominalization (also called nominalization) is the use of a word

form which actually does not exist as a noun but is being used as a noun or as the head of a noun phrase, with or without morphological transformation.

The proposed models of lemmatization – namely LemmaChase and LemmaQuest, both successfully extract lemma for all allied morphed words like nouns, verbs, derivative words and especially nominalized words. The LemmaChase lemmatizer generates 85% - 90% correct lemmas for individual morphed input words and covers all nominalized words and single-suffixed derived words accurately. As compared to this the existing lemmatizers do not generate the lemmas for derived and nominalized words at all.

The second model LemmaQuest is a superset of LemmaChase and it generates lemmas after creating groups of related allied words. The design of LemmaQuest is based on a combination of language-independent statistical distance measures, segmentation technique, rule-based stemming approach and lastly morphological transducer's parsing technique to generate the correct dictionary word for a set of related morphed words. Through a single pass, the proposed model completes lemma generation process which leads to improve execution rate of the lemmatization process and correctly handles morphological nominalized words. LemmaQuest is able to handle more variety of derived words like double and triple suffixed words and is able generate 90% correct lemmas.

The output of both the models has been compared with existing popular lemmatizers showing significant advanced lemma generation especially for nominalized words using the standard datasets available in the DUC-text corpus, Brown corpus, the Oxford Dictionary and the MorphoLex repository. As compared to these two models, most of the available popular lemmatizers fail to generate correct lemmas for any of the derived and nominalized words. The research work conducted has been published in UGC Care Journal, IETE Journal and Springer Publication.