Chapter 1

Introduction to Lemmatization

1.1 Introduction

This chapter discusses the research gap, problem formulation, problem statement, research objectives and research contribution of the work done and research conducted. Over here, the discussion is based on the morphological structure of English words which eventually leads to the designing and developing of the proposed models. It also covers morphological analysis, stemming and lemmatization applied on a word.

To design the objectives and to formulate the problem statement it is important to understand in detail the notion of morphology, morphemes, lexemes and working of a morphological analyzer. In this chapter therefore the focus is on understanding these concepts.

1.2 Morphology

To reach the problem statement and understand and comprehend the need and importance of the work done in this research, it is vital to understand certain very important concepts related to morphology. The initial part of this chapter therefore focuses on the discussion of the various terms related to morphology and how significant it is to the ultimate goal of designing and developing a lemmatizer. “The term **‘word’** is part of everyone’s vocabulary. The area of grammar concerned with the structure of words and with relationships between words, is technically called **morphology**” (Andrew Carstairs-McCarthy, 2002). It discusses the ways in which new and complex words are formed. Morphology is also called ‘The Poland of Linguistics’ (Spencer and Zwicky, 1998), surrounded by neighboring fields like Text Mining, NLP, Machine Translation etc which are eager to claim the territory for themselves. It is responsible for describing the internal structure of any complex words.

“The term **morphology** is generally attributed to the German poet, novelist, playwright, and philosopher Johann Wolfgang von Goethe (1749–1832), who coined it early in the nineteenth century in a biological context. Its etymology is Greek: morph means ‘shape or form’, and morphology is the study of form or forms. In biology, morphology refers to the study of the form and structure of organisms. In linguistics, morphology refers to the mental system involved in word formation or to the branch of linguistics that deals with words, their internal structure, and how they are formed.” (Mark Aronoff and Kirsten Fudeman, 2010)

1.3 History of Morphology

In early age, traditional linguistic analysis had treated the **word** as the basic unit of grammatical theory and lexicography. In the year between 1920 and 1945, American structuralists grappled with the problem of how sounds are used to distinguish meaning in language. They developed and refined the theory of the phoneme (Bloomfield, 1933). When structuralism was in its prime, especially between 1940 and 1960, the study of morphology occupied centre stage. Many major structuralists investigated issues in the theory of word-structure. Eugene A. Nida’s course book titled Morphology, which was published in 1949, codified structuralism theory and practice. (Eugene A. Nida, 1949)

Leonard Bloomfield (1887-1949), together with Edward SAPIR, was two most prominent American linguists of the first half of the twentieth century who explored morphological structure of the words (Bloomfield, 1933). Bloomfield with his students, particularly Bernard Bloch, Zellig Harris, and Charles Hockett established the school of thought that has come to be known as American structural linguistics, which dominated the field until the rise of Generative Grammar in the 1960s (Sapir, 1925; Swadesh, 1934; Twaddell, 1935; Harris, 1944).

Zellig Harris (1955) developed some indirect procedure to segment the words into morpheme based on their utterance. In later stage, Hafer’s & Weiss’s Word Segmentation methods (1974) based on Harris' segmentation concept had become popular to generate stem from a word. The linguistics’ concept[1] of any English word, are layered into various levels which are shown in Fig. 1.1.

[1] https://www.britannica.com/science/linguistics/Morphology

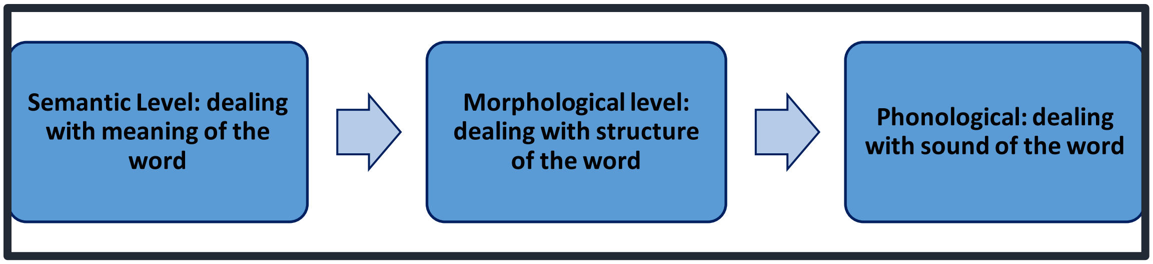
Semantic Level: dealing with meaning of the word
Morphological level: dealing with structure of the word
Phonological: dealing with sound of the word

Figure 1.1 Levels of linguistic concept

1.4 Morpheme

Morphology is the study of morpheme which is described as the smallest meaningful linguistic piece with a grammatical function. The analysis of words into morphemes begins with the isolation of morphs which is the physical form representing some morpheme in a language. It is a recurrent distinctive sound (phoneme) or sequence of sounds (phonemes).

Two broad classes of morphemes can be distinguished: stems—the central morpheme of the word, supplying the main meaning and affixes—adding supplementary meanings of various kinds. Affix morphemes can be divided into two major functional categories, namely derivational morphemes and inflectional morphemes. This reflects recognition of two principal word building processes: inflection and derivation. Inflectional and derivational morphemes form words in different ways. Sometimes the presence of a derivational affix causes a major grammatical change, involving moving the base from one word-class into another as in the case of -less which turns a noun into an adjective. In case of derivational words, e.g. the word “relational” is morphologically analyzed into the central morpheme “relate” with association of suffix “-ational” [-ational → -ate.]; word “organization” should be analyzed into “organize” and “-ation”, not into “organ” and” -ization”; word “policy” should be identified as a central morpheme, should not be converted into “police” (Daniel Jurafsky and James H. Martin, 2020). All these above mentioned complex English words should be analyzed accurately through morphological analyzer which presently not been done by any existing lemmatizer.

To realize the importance of designing a lemmatizer, it is important to understand the process of construction of any complex word* along *with understanding the splitting of morphemes.

Table 1.1 shows a sample list of derivation suffix which are attached with some words to form derivative complex words.

Table 1.1 Derivational Morpheme

Input Base	Derivational Suffix	Word-class of input base	Word-class of output base	Output Base
child	-hood	Noun	Noun	child-hood
king	ship	Noun	Noun	king-ship
kind	-ness	Adjective	Noun	kind-ness
kind	-ly	Adjective	Adverb	kind-ly
sincere	-ity	Adjective	Noun	sincer-ity
govern	-ment	Verb	Noun	govern-ment
power	-less	Noun	Adjective	power-less
power	-ful	Noun	Adjective	power-ful
democrat	-ic	Noun	Adjective	democrat-ic
refuse	-a1	Verb	Noun (abstract)	refus-al
read	-er	Verb	Noun	read-er
anarchy	-ist	Noun	Noun	anarch-ist
piano	-ist	Noun	Noun	pian-ist

As shown in Table 1.2, the diminutive suffix -let (Seq. 1a) is attached to nouns to form diminutive nouns (meaning a small something). The derivational suffix -ship (Seq. 1b) is used to change a concrete noun base into an abstract noun (meaning ‘state, condition’).

Table 1.2 Derivational Morpheme for Noun

Seq.	Input Base in NOUN class	Output Based on Diminutive Noun /Abstract Noun class
1.a	pig	pig-let
1.a	book	book-let
1.b	friend	friend-ship
1.b	leader	leader-ship

The morphemes carry morphological, phonological, syntactical, semantical and functional information. The nature of the morpheme is depicted in Fig. 1.2. It contains morphological information, phonological information, syntactic, semantic and functional information.

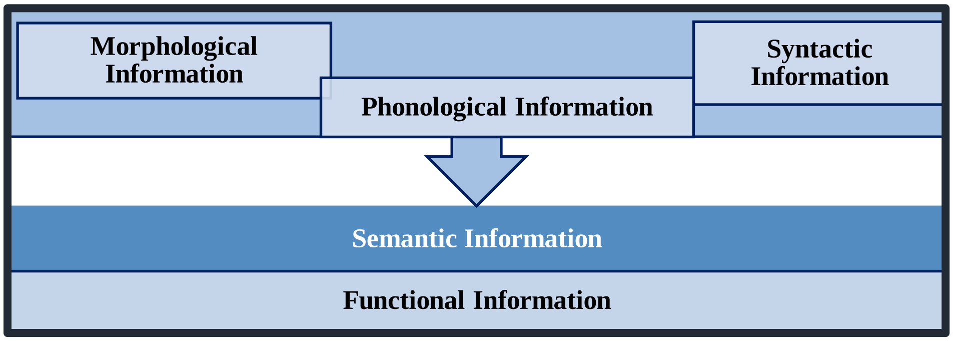

Figure 1.2 The Nature of Morpheme

An important concept in grammar and more particularly, in morphology is that of free and bound forms. A bound form is one that cannot occur alone as a complete utterance (in some normal context of use). For example, -ing is bound in this sense, whereas wait is not, nor is waiting. Any form that is not bound is free. Any free form that was not a phrase was defined to be a word and to fall within the scope of morphology. One of the consequences of Bloomfield’s definition of the word was that morphology became the study of constructions involving bound forms. The so-called isolating languages, which make no use of bound forms (e.g. Vietnamese) would have no morphology.

Types of Morphemes is shown in Fig. 1.3 which depicts different forms of morphemes.

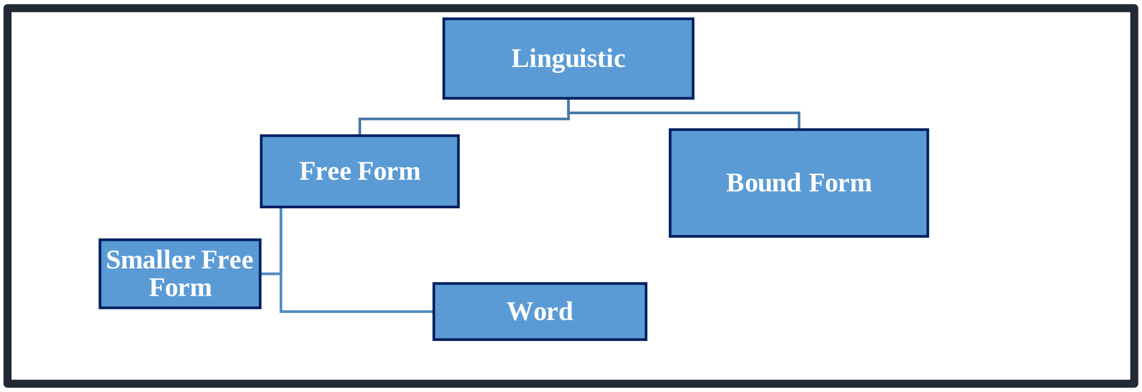

Figure 1.3 Types of Morphemes

|

1.5 Types of Words

Linguists distinguish the word into three different categories: phonological words, grammatical words and lexemes which is shown in Fig. 1.4.

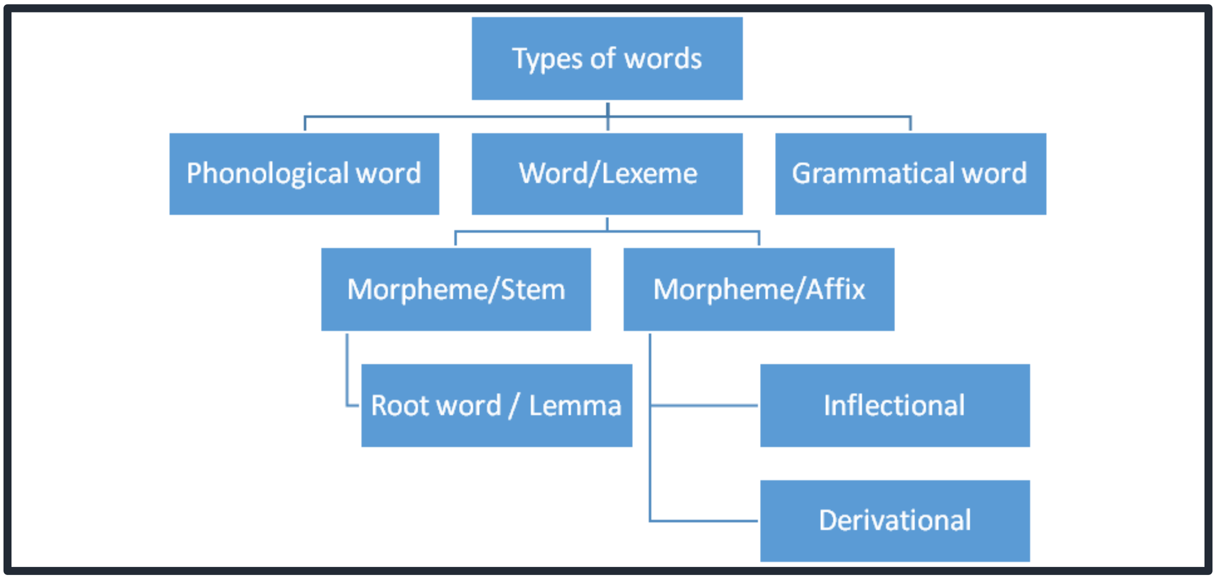

Figure 1.4 Types of Words

A **phonological word** can be defined as a string of sounds that behaves as a unit for certain kinds of phonological processes, especially stress or accent.

The term **grammatical word** or morpho-syntactic word is virtually synonymous with word but is generally used to refer specifically to different forms of a single word that occur depending on the syntactic context. For example, woman and women are tokens of the same word, but they absolutely must be considered to be different grammatical words. The first occurs in context for a singular noun, and the second in context for a plural noun. Even though forms “like”, “and” and “into” have only one form, they are also considered grammatical words.

Lexeme:

A **lexeme** is a word with a specific sound and a specific meaning. Its shape may vary depending on syntactic context. A lexeme is a theoretical construct that corresponds roughly to one of the common senses of the term word. A lexeme is a unit of morphological analysis which represents a set of morphed words in linguistics. Two types of lexemes are available for the formation of word structure: inflection and derivation in English grammar. Inflection involves the formation of grammatical forms: past, present, future; singular, plural; masculine, feminine and neuter. The use of these grammatical forms is generally ruled by sentence structure. In English, regular verb lexemes have a lexical stem, which is the bare form with no affixes (e.g., jump) and three more inflected forms, one each with the suffixes -s, -ed, and -ing (jump, jumped, and jumping). Noun lexemes have a singular and plural form. Adjectives, adverbs, prepositions, and other parts of speech typically exist in a single form in English (Bloomfield, 1939).

Lexeme in derivation involves the creation of a new lexeme from an existing lexeme, such as “employer” or “employment” from “employ”. In many cases, the Part-Of-Speech (POS) of a new lexeme has been changed into different POS’s class after attaching a derivational suffix. E.g. Verbs to nouns: employ + er; Verb to Adjective: accept + able; Nouns to nouns: fish + ery; India + ian; Adjectives to adjectives: blue + ish; e.g. Derivative function creates agent nouns from verbs : $X]_V$ er$]_N$; e.g.: think$]_V$ er$]_N$, run$\underline{n}]_V$ er$]_N$ etc. (Mark Aronoff ; 2010)

In dictionary, there is no logical or syntactical mapping between derivative lexeme and root lexeme; both exist as independent words in any dictionary. So there was a need to design a model which can extract root lexeme from derivational lexeme.

Words can also be categorized into Content words and Function words. Finegan (1994: 161) expressed the difference well and he wrote that content words “have meaning in that they refer to objects, events, and abstract concepts; are marked as being characteristic of particular social, ethnic, and regional dialects and of particular contexts; and convey information about the feelings and attitudes of language users. Function words also have meaning, but in a different way”.

To summarize, a lexeme is an abstract object, not a single concrete word, but a set of grammatical words. Cross-linguistically, one of those words is generally privileged to be the lexical stem from which other words are formed, although some languages permit more than one lexical stem. After understanding the difference between words and lexemes, it is easy to make distinction made by morphologists between inflection and derivation.

Lexeme with Multiple Affixations

Sometimes, it is observed that complex words/lexemes are formed by creating bases which contain several derivational morphemes. This process can take place in a number of rounds, with the output created by one round of affixation serving as the input to a later round. Sample lexemes with double suffixation are shown in Table 1.3.

Table 1.3 Double Suffixation words

Seq. No	Input Lexeme	Single Suffix Deletion Lexeme	Double Suffix Deleted Lemma/Root Word
1.	developmental	development (Derivative)	develop
2.	applicability	applicable (Derivative)	apply
3.	applicants	applicant (Inflected)	apply
4.	applications	application (Inflected)	apply
5.	appliances	appliance (Inflected)	apply
6.	complaints	complaint (Inflected)	complain
7.	complications	complication (Inflected)	complex
8.	compliances	compliance (Inflected)	comply
9.	privatization	privatize (Derivative)	private
10.	administrations	administration (Inflected)	administrate
11.	deployments	deployment (Inflected)	deploy
12.	*deployable	*Not available in some dictionaries (in WordNet)	deploy
13.	employees	employee (Inflected)	employ
14.	suppliers	supplier (Inflected)	supply

Roots: A root is the irreducible core of a word, with absolutely nothing else attached to it. It is the part that is always present, possibly with some modification, in the various manifestations of a lexeme. For example, walk is a root and it appears in the set of word-forms that instantiate the lexeme “walk” such as walk, walks, walking and walked, walker. Many words contain a root standing on its own. Roots which are capable of standing independently are called free morphemes, e.g. man, book, tea, sweet etc.

Many other free morphemes are function words. These differ from lexical morphemes in that while the lexical morphemes carry most of the ‘semantic content’, the function words mainly (but not exclusively) signal grammatical information or logical relations in a sentence. Typical function words include the following:

Articles: a, the .

Demonstratives: this, that, these, those.

Pronouns: I, you, we, they, them; my, your, his, hers; who, whom, which, whose, etc.

Conjunctions: and, yet, if, but, however, or, etc.

So, at the time of designing and developing a lemmatization model, all above mentioned function words are not considered for processing because all these words are not carrying any special meaning independently.

1.6 Approaches of Morphological Analyzer

There are two complementary approaches to morphology, **analytic and synthetic.** The analytic approach has to do with breaking words down. Some basic analytic principles used in morphology, are described in textbook “Morphology” by Eugene Nida’s (1949; revised edition 1965), mentioned earlier. This book describes six principles employed in isolation and in identification of Morphemes. To isolate and to reconstruct the morphemes from words, are the major challenging task on which NO WORK has been done. None of the principles of Eugene Nida, is complete in itself; each is supplementary to the basic definition and must be considered so. The first principle is stated as: Forms which have a common semantic distinctiveness and an identical phonemic form in all their occurrences constitute a single morpheme (Eugene Nida, 1949).

Principle 1: It means that such a form as morph “**-er”** added to verbs in such constructions as work-er, dance-er, runn-er, walk-er, and fli-er is a morpheme. It always has the same phonetic form, and always has essentially the same meaning, namely, that of 'the doer of the action' (also called 'agentive'). The principle used the phrase "common semantic distinctiveness" as a way of indicating that the meaning which is in common to all the occurrences of the suffix “-er” contrasts with the meaning of all other similar forms. In the definition of the morpheme and in the statement of this first principle, it was observed that the meaning of “-er” in all these positions is not necessarily identical. In fact, no science has made available to us the tools by which the degrees of difference in meaning can be tested. For example, in English there is another suffixed morpheme with the form “-er”, that is, the “-er” in comparative adjectives such as wid-er, broad-er, small-er, deep-er, clean-er. There is however no common semantic distinctiveness in the series of suffixed forms occurring in work-er, danc-er, runn-er.

Principle 2: Forms with the same meaning but different sound shapes may be instances of the same morpheme if their distributions do not overlap.

Principle 3: Not all morphemes are segmental e.g. breath$_N$ breathe$_V$; cloth$_N$ clothe$_V$.

Subscript “N” for Noun, “V” for Verb.

Principle 4: A morpheme may have zero as one of its allomorphs (an allomorph is a variant phonetic form of a morpheme. e.g. buses /bʌsəz/) provided it has a non-zero allomorph (any of two or more actual representations of a morpheme). Fish generally displays no special marking in the plural: one fish, ten fish-Ø (Ø indicates no suffix).

So, for isolation of morphemes from a word, the type of words should be first identified.

1.7 Application of Morphological Analyzer

Morphological analyzers are executed as a prerequisite task for any Information Retrieval (IR) or Information Extraction (IE) applications to reduce the number surface words of any text to their corresponding root words. Surface words are mapped to limited number of morphemes or roots words through morphological analyzer.

Currently search engine’s functions are limited to retrieving the URLs of only the input words. However, with the assistance of the designed and developed lemmatization models, a search engine would be able to retrieve URLs related to input words as well as their intermediate words and root words.

The Machine translation systems need to analyze words to their components through morphological analyzer and generate words with specific features in the target language. The computational morphology extracts any information encoded in a word and brings it out so that later layers of IE processing can make use of it [employee/employer → employ]. Under computational process, derivational morphology construct a new word with usually a different part-of-speech category which is still unexplored by existing lemmatizes and also by existing morphological analyzers [e.g. nation/national/nationalize/nationalist/nationalism]. The existing lemmatizers and morphological analyzers can only reduce inflected surface words into its lemma or root word. Morphological Analyzers as popularly developed are also called Stemmers and lemmatizers. Both of them process the words to convert them in normalized form. **Normalization** is a process to map allied morphed words to a single root word.

1.8 History of Application of Morphological Analysis

The text written in natural language was started to be analyzed automatically from 1954. The historical description of morphological analysis is shown in Table 1.4.

Table 1.4 History of IE, IR and NLP

IE:	
1954	The first automatic translations tool had been developed from English to the Russian language in 1954, but automation was limited to a handful of sentences. [2]
late 1970s	Automation in "Information Extraction" had been initiated.
Mid-1980	JASPER tool had been popularized for Reuters to provide real-time financial news.
Beginning in 1987, 1989	IE was spurred by a series of Message Understanding Conferences. MUC is a competition-based conference that focused on the following domains: MUC-1 (1987), MUC-2 (1989): Naval operations messages.
1998	MUC-7: Satellite launches reports.
IR:	
1970s	First online systems—NLM's AIM-TWX, MEDLINE; Lockheed's Dialog; SDC's ORBIT.[3]
1975	Three highly influential publications by Salton fully articulated his vector processing framework and term discrimination model. A Theory of Term Importance in Automatic Text Analysis (JASIS v. 26).
1979	C. J. van Rijsbergen published Information Retrieval (Butterworths). Heavy emphasis on probabilistic models.
1989	First World Wide Web proposals by Tim Berners-Lee at CERN.
1999	Publication of Ricardo Baeza-Yates and Berthier Ribeiro-Neto's Modern Information Retrieval by Addison Wesley, the first book that attempts to cover all IR.
NLP:	
1960	Some notably successful natural language processing systems that were developed in the 1960s were SHRDLU.[4]
1966	ELIZA, a simulation of a Rogerian psychotherapist, written by Joseph Weizenbaum was created.

[2] https://en.wikipedia.org/wiki/Information_extraction

[3] https://en.wikipedia.org/wiki/Information_retrieval

[4] https://en.wikipedia.org/wiki/Natural_language_processing

1990	The research on the core topics such as word sense disambiguation and statistically colored NLP got a direction of research.
1990s– 2010	**Statistical NLP**: Up to the 1980s, most NLP systems were based on complex sets of hand-written rules. Starting in the late 1980s, there was a revolution in NLP with the introduction of machine learning algorithms for language processing. With the growth of the web, research has thus increasingly focused on unsupervised and semi-supervised learning algorithms. e.g., text and speech processing, text-to-speech processing, word segmentation (tokenization), lemmatization and stemming, morphological segmentation and part-of-speech tagging.

1.9 Stemming and Lemmatization

Stemming reduces morphological variants into a common stem and lemmatization process reduces to dictionary root word or lemma. Both are doing morphological analysis of any type of morphed words to get or to construct root stem or lemma. The objective of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For grammatical reasons, different morphed words exist in the documents, such as organize, organizes, organizing, organizer and organization; democracy, democratic and democratization; etc. In many situations, it seems as if it would be effective and beneficial for a search for one of these words to return documents that contain another word in the set.

In stemming and lemmatization processes, two term conflation approaches: non-linguistic and linguistic are basically implemented. Most of the Stemming algorithms are designed based on linguistic studies and on word morphology; they do not utilize methods pertaining to NLP. Fig. 1.5 depicts a concise categorization of term conflation methods (Carmen Galvez, Fe´lix de Moya-Anego´n and Victor H. Solana, 2005. pp. 520-547).

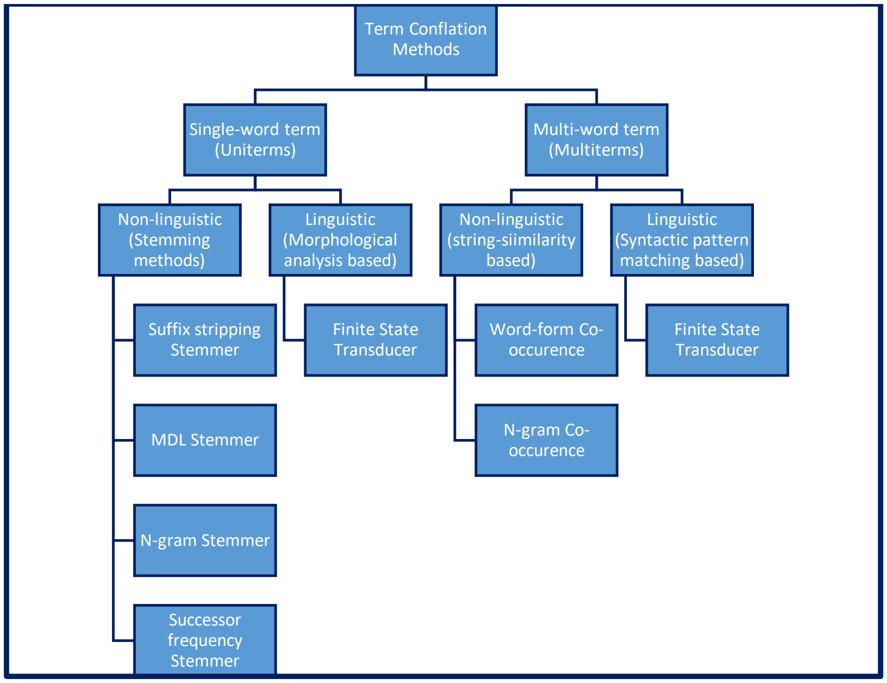
Term Conflation Methods
Single-word term (Uniterms)
Multi-word term (Multiterms)
Non-linguistic (Stemming methods)
Linguistic (Morphological analysis based)
Non-linguistic (string-siimilarity based)
Linguistic (Syntactic pattern matching based)
Suffix stripping Stemmer
Finite State Transducer
Word-form Co-occurence
Finite State Transducer
MDL Stemmer
N-gram Co-occurence
N-gram Stemmer
Successor frequency Stemmer

Figure 1.5 Term Conflation Methods

Now, in this research work, "Term-Conflation-Methods" are applied only on single-word term (Uniterms). Many stemming algorithms generate good results for the conflation and normalization of uni-term variants (Porter, 1980; Frakes, 1992). Within suffix-stripping stemmer, the most effective technique is the **longest match algorithm** which was applied in Lovins' stemming procedure (J. B. Lovins, 1968). In stemming, the morpheme left after affix elimination can hardly be used for IR and IE purposes because most of the time, that morpheme may not be a dictionary word. The solution to this problem resides in doing a full morphological analysis, and this task can only be processed by Lemmatizer (Carmen Galvez, 2005).

In an article, Goldsmith (2001) claimed that automatic morphological analysis can be divided into four major approaches. The first approach, based on the work explored by Harris (1955) and further designed by Hafer and Weiss (1974), provides a goodness of break between letters that can be compute by the successor frequency there, compared to the successor frequency of the letters on either side. The second approach looks for n-grams, which are likely to be morpheme-internal (Adamson and Boreham, 1974). The third approach highlights the discovery of patterns of phonological and morphological relationships between pairs of words: a base form and an inflected form. The fourth approach explores unsupervised learning techniques, yielding a partition of stems and affixes, segmenting longer strings into smaller units (Kazakov, 1997; Kazakov and Manandhar, 2001).

(1) **Non-linguistic Approach**: "**Stemming** methods consist mainly of suffix stripping, stem-suffix segmentation rules, similarity measures and clustering techniques." (Carmen Galvez, Fe´lix de Moya-Anego´n and Victor H. Solana, 2005)

(2) **Linguistic Approach**: "**Lemmatization** methods consist of morphological analysis. That is, term conflation based on the regular relations, or equivalence relations, between inflectional forms and canonical forms, represented in finite-state transducers (FST)." (Carmen Galvez)

Some studies have uncovered that indexing by the stem does not substantially improve the performance of retrieval, at least not in the English language (Harman, 1991)[5] (W. B. Frakes, Carmen Galvez). Researcher, Harman, observed that the use of a stemmer in the query is intuitive to many users, and reduces the number of terms decreasing the size of the index files, but retrieves too many non-relevant documents. This problem can be minimized by the process of lemmatization (Carmen Galvez, 2005). The proposed lemmatization model is developed based on both stemming and lemmatization techniques.

1.10 Applications of Stemming and Lemmatization

Most useful applications of IE, IR and NLP cannot be developed without the implementation of stemming or lemmatization. Some are described below.

i) Text-Summarization: In Automatic Text-Summarization, preprocessing is an important phase to reduce the space of textual representation. Classically, stemming and lemmatization have been widely used for normalizing words. However, even using normalization on large texts, the curse of dimensionality can disturb the performance of summarizers (Juan-Manuel Torres-Moreno, 2012). Morphological analysis is a very important phase of pre-processing of NLP systems because it allows reducing the dimension of the vector space representation in an IR system (R. Baeza-Yates and B. Ribeiro-Neto, 1999; C.D. Manning and H. Schütze, 1999). Several other applications such as Document Indexing, Textual Classification and Question-Answering systems among others utilize this reduction.

[5] https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

ii) Application of NLP: Regardless of our text data format, steps that are used to solve NLP problems remain more or less same. Major steps are depicted in Fig 1.6 and Fig. 1.7, while solving the NLP problems.

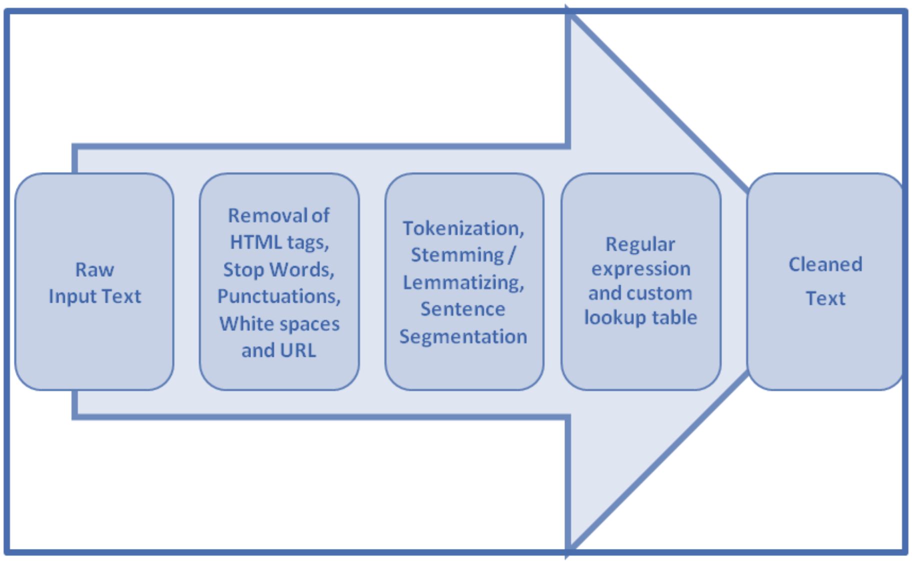
Raw Input Text
Removal of HTML tags, Stop Words, Punctuations, White spaces and URL
Tokenization, Stemming / Lemmatizing, Sentence Segmentation
Regular expression and custom lookup table
Cleaned Text

Figure 1.6 Preprocessing Steps in NLP

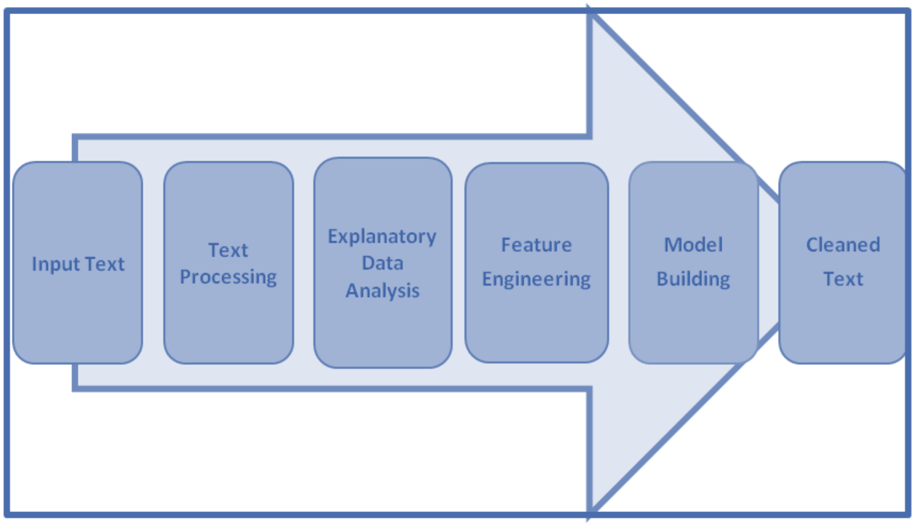

Figure 1.7 Steps in NLP

iii) Text Normalization/Text Simplification: On a higher level, normalization is used to reduce the dimensions of the features, so some machine learning models can efficiently process the data. Text data contains multiple representations of the same word which will be merged into a single word through lemmatization. e.g., “player”, “played”, “plays” and “playing” are different variations of the word “play”.

iv) Information Retrieval Model: It describes a computational model for representing documents and queries as well as how the two representations are matched (Manning, Raghavan, and Schütze, 2008). The representations are used to compute a measure of similarity between a query and a document. Similarity measures are usually based on stemming or lemmatization process and similarity distance measure between pair of strings. IR model checks the frequency of the terms in a query has been exist in a document, i.e., assuming topicality sense of relevance. Information retrieval systems use retrieval models to match and rank documents.

1.11 Research Gap and Research Formulation

One of the important pre-requisite tasks of TM, IE, IR and NLP applications like Name Entity Recognition, Summarization, Text Simplification, Semantic Analysis, etc. is to trim the text into number of significant base words through which sentiment of the text would be accurately mined. For extracting a limited collection of dictionary base-words from all morphed and derived words available in the given text, lemmatization process is more effective as compared to stemming process because stemming processes are not able to generate dictionary base-words most of the time.

It has been observed that the existing popular lemmatizers are still not handling the derived morphed words and nominalized words like application, applicant, judgmental, employer, employee, safely, etc., in the sense that the lemmas of these words are not correctly generated. They can only handle inflected words like women, running, broken, applying, applied, employs, etc. to generate their lemmas. Some of the stemming processes can properly handle both inflected and derived words and can generate single stem for all morphed allied words e.g. the Porter’s Stemmer generates single stem ‘applic’ for the words – applicant, application and ‘accept’ for the words – acceptance, acceptable. However, most of the time, all these stems are not necessarily dictionary words like the stem ‘applic’ which is not a real word.

Like stemmers, most of the popular lemmatizers are also not able to generate the correct lemmas for all English derived words. Therefore only stemming or only lemmatization process cannot provide us a proper solution in generating the dictionary root words for both inflected and derived English words. This limitation can affect the NLP, TM, IE, etc. applications and their output drastically. Hence it was observed after detailed literature survey that the research gap was there in the preprocessing part of any Text Mining and NLP related applications. Consequently a need to develop a lemmatizer which would handle this research gap and generate a better pre-processing output was realized.

A possible solution would be to generate stem token through suitable stemming process and then to construct a meaningful dictionary word from stem token through proper morphological analysis. The research formulation drives towards designing and developing a simpler and efficient lemmatization model which can generate proper lemmas with minimum error and a very competitive output as compared to the lemmas generated by the currently prevalent lemmatizers.

1.12 Problem Statement

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.

1.13 Objectives of the Study

The objective of this research work is to develop a simple, robust and enhanced model for lemmatization. This model is proposed to overcome the shortcoming of the existing stemmers and lemmatizers, generating an accurate, precise and exact output in terms of lemmas for the input text. The work also focuses on suffixes and not prefixes of the morphed words as handled by any lemmatizer or morphological analyzer. This intended research work has the following objectives:

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

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.
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.
4. The lemmatizer should work on the text as a whole combining related morphed /derived words making the processing and output simpler and acceptable.

1.14 Research Contribution

Two models are presented and implemented that offer a generic and logical solution to meet the research objectives.

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

The second proposed model-LemmaQuest is very effective 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 can generate a single lemma for a group instead of generating a lemma for an individual input word. 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 precise, and the error rate is far less as compared to the output by other prevailing lemmatizers.

1.15 Organization of the thesis

The contributions from this study have been presented in the chapters as follows:

Chapter-1: This chapter contains the general introduction to morphology, stemming and lemmatization. The limitations of existing lemmatizers which inspires for the research problem statement development and objectives of the research contribution have been discussed here.

Chapter-2: An extensive literature study conducted on different types of stemming and lemmatizations methods has been carried out and presented here. The discussion on different distance measures related to unsupervised approaches of stemming and morphological analysis is also discussed here.

Chapter-3: As a conclusion to the previous literature study, this chapter focuses on the implementations of different stemmers and lemmatizers. The detailed comparative analysis has been done after implementing various stemmers and lemmatizers.

Chapter-4: This chapter discusses in detail the lemmatizer model designed and developed as part of the research conducted which handles the challenging limitations of existing lemmatizers. It has been named as LemmaChase. The design and implementation of LemmaChase with comparisons is deliberated upon here.

Chapter-5: This Chapter focuses on a novel lemmatizer which implements statistical distance measure to make the previous model more efficient and handles nominalization more accurately. This model has been named LemmaQuest. The LemmaQuest model design, implementation, comparison and its assessment with LemmaChase is explained in detail in this chapter.

Chapter-6: Conclusion and Future work.

This is followed by Publication detail and References.

1.16 Conclusion and Summary

In this chapter, the discussion was based on morphology and overview of stemming / lemmatization. The limitations of existing popular lemmatizers inspired to formulate the problem statement and to conduct the research work.