Chapter 4

LemmaChase: A Lemmatizer

4.1 Introduction

In this chapter, the discussion is based on the proposed lemmatizer: LemmaChase which is the first model designed and developed as part of the research work conducted. This model was designed to eliminate the shortcomings of the currently available popular Lemmatizers like the Stanford LemmaProcessor, spaCy Lemmatizer, LemmaGen and morphological analyzer like MorphAdorner etc. This model takes into account the nominalized/derived words for which correct lemmas are currently not generated by any available lemmatizer or morphological analyzer.

4.2 Understanding of LemmaChase

To develop a lemmatizer, the foremost challenge lies in understanding the morphological structure of any input English word and especially comprehending the derivational word's structure. Another important challenging task for a lemmatizer is identification of derivational suffix from morphed words and then extraction of dictionary base word from that derived word. Existing famous and popular lemmatizers are not handling such derivative words to extract their base words. The proposed lemmatizer – LemmaChase, extracts the base word correctly considering the pre-requisite knowledge of the word's Part of Speech(POS), different class of suffix rules and efficiently implementing the recoding rules. LemmaChase successfully generates the base word form from all its derivational / nominalized word forms available in any standard English dictionary like Oxford and Cambridge.

The broad view of the LemmaChase lemmatizer is depicted in the flowcharts given below. Figure 4.1 depicts the overview of the model. The preprocessing required before lemmatizer is implemented is shown in Figure 4.2. The next step involves the categorization and parsing of each input word which is depicted in Figure 4.3. Figure 4.4 shows the final flow of the LemmaChase lemmatizer with lemmas which are either extracted or generated.

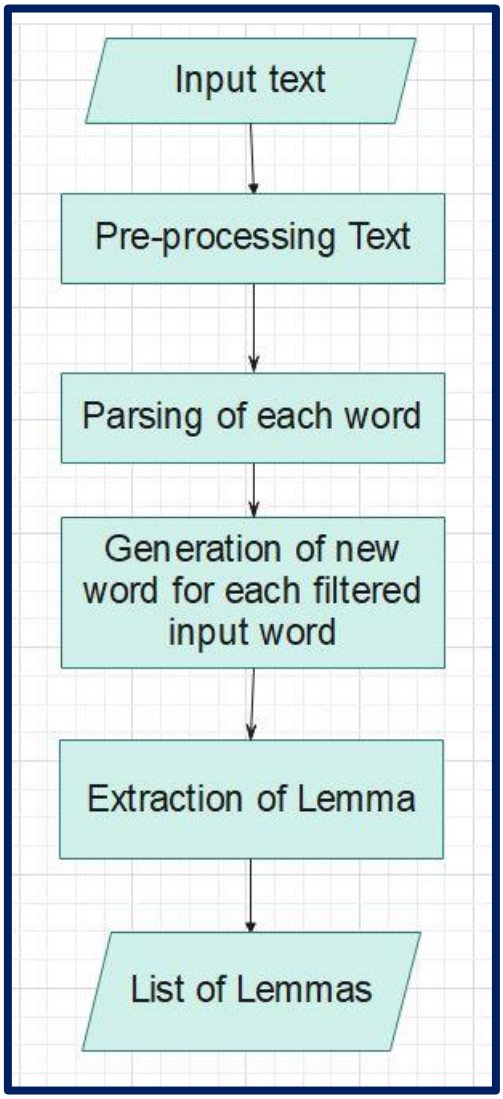


Figure 4.1 Overview of LemmaChase

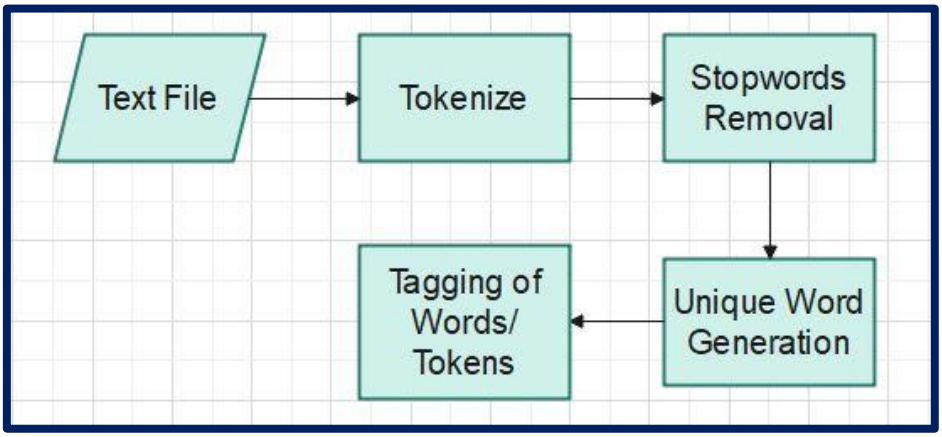


Figure 4.2 Pre-Processing of Text

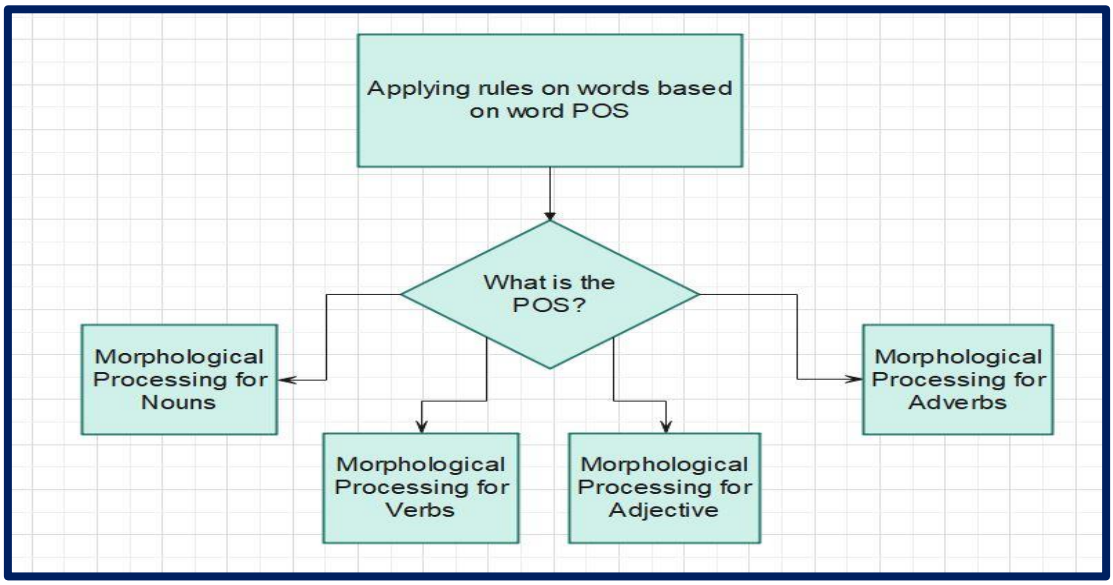


Figure 4.3 Parsing of Words

Note: POS (part-of-speech)

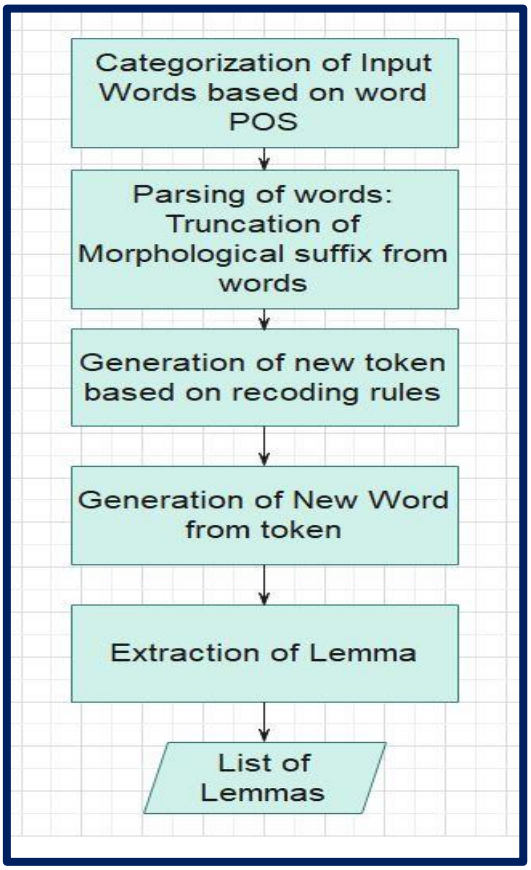


Figure 4.4 Extraction/Generation of Lemma

The LemmaChase lemmatizer and the algorithm behind it, is described in detail in the next section.

4.3 Input DataSet

The input datasets are taken from a number of standard sources. The list of words is constructed from online Oxford dictionary[1], Merriam-Webster.com online dictionary[2] and MorphoLexDatabase[3].

Different sets of input text are also taken from Documents Understanding Conferences (DUC-2004) Corpus[4] and Brown Corpus[5].

DUC published text corpus of 500 news articles for text summarization. Specifically it consists of 50 clusters of Text Retrieval Conference (TREC) documents, from the following collections: AP newswire, 1998-2000; New York Times newswire, 1998-2000; Xinhua News Agency (English version), 1996-2000. Each cluster contained on average 10 documents (DUC Corpus).

The Brown University Standard Corpus of American English (Brown Corpus) is an electronic collection of text samples, the first major structured corpus of varied genres. This corpus was compiled and published by Henry Kučera and W. Nelson Francis at Brown University, in Rhode Island. It is a general language corpus containing 500 samples of English, totaling roughly one million words.

Claudia H Sánchez-Gutiérrez, Hugo Mailhot and S Hélène Deacon and Maximiliano A Wilson published sizeable and freely available database 68,624 morphologically complex words (MorphoLexDatabase: "A derivational morphological database by Claudia H Sánchez-Gutiérrez, Hugo Mailhot, S Hélène Deacon and Maximiliano A Wilson") which were constructed from the English Lexicon Project. This database represents a valuable resource for studies on the effect of morphology in visual word processing in English.

These are the datasets being popularly used for all Text Mining, NLP and Computational Linguistic related research work.

[1] https://www.oxfordlearnersdictionaries.com/wordlist/american_english/oxford3000/
[2] https://www.merriam-webster.com/browse/thesaurus/
[3] https://github.com/hugomailhot/MorphoLex-en/blob/master/MorphoLEX_en.xlsx
[4] https://duc.nist.gov/data.html
[5] https://www.kaggle.com/datasets/nltkdata/brown-corpus

4.4 Algorithm of LemmaChase

Before the algorithm is discussed, it is important to understand the standard preprocessing which is normally done on the text file for any text mining related applications. Steps of the preprocessing are as follows.

Preprocessing of the File:

1. In this step, all unnecessary characters like punctuations, symbols white space will be removed.
2. Convert all words into lower case and stop words are removed.
3. Tokenize the sentence into words.
4. Generate sorted unique word list.
5. Tag all unique words using Stanford POS tagger.

Details of LemmaChase Lemmatizer:

Input: List [w_s]={w_i/w_i_POS_tag, w_j/w_j_POS_tag…w_n/.. }: w_i represents i^{th} word of the text. w_i_POS_tag represents the context (Part-Of-Speech) of word w_i in the text.

Process: Each surface word (w_s/w_i) to be lemmatized through invocation of function**:** **"$f_{LemmaChaset}$(List[w_s])".**

Output: S_L: List [lemmas]. S_L represents a compressive list of lemmas generated for the input-words'.

S_L = {L_1, L_2, L_3.....}

Note: For any type of English derived words, WordNet dictionary is not able to extract their root word or lemma. Dictionary shows the input derived word to be the lemma itself. (E.g. for the words like "deployment", "employability", the lemmas are also "deployment" and "employability")

Step: 1–Parsing of Proper Noun and Verb POS tagged words.

Based on token's POS, words are parsed with the help of morphological rules as given in Table 4.7 to get their corresponding Verb lemma from Verb tagged word. Elaborated parsing explanation is given in Section 4.5.

- Proper Noun tagged word is not parsed further, same input word will be identified as its lemma. e.g., word "India" is identified as lemma "India". "Indian" is identified as a Noun and an Adjective both which would be parsed further.
- If a word's POS is identified only as Verb in any tense in WordNet dictionary using JWNL (Java WordNet Libraries) API, then the lemma extracted for that word is

finalized as the final lemma. e.g., lemma "achieve", "go" and "program" are finalized as the lemma for the words "achieved", "went" and "programmed".

If the verb (e.g. $signify_{Verb} \rightarrow$ sign, $carbonify_{Verb} \rightarrow$ carbon, $historify_{Verb} \rightarrow$ history)[6] is identified as a nominalized word, then word-parsing is required to generate new word in Verb form or in Noun form.

Step: 2 –Parsing of words, tagged both in Common Noun and Verb POS / both in Adjective and Verb POS.

- If the input word is identified as both Noun/ Adjective and Verb POS form, lemma of Verb form's word is selected as lemma if and only if length of lemma <= length of input word. e.g., "programming", "photograph", "influence" and "left" exist as both Verb and Noun. The word "left" is processed further.

Step: 3 –Parsing of word, tagged only in Common Noun or only in Adjective POS.

- If the input word only exists as a Noun or as an Adjective in WordNet dictionary and its extracted lemma is the same as that of the word itself, then such observation indicates that this lemma could be correct or incorrect for that input word. For nominalized noun/ derived adjective word (e.g. application, applicant, employee, employer, applicable, productive), WordNet dictionary is not able to extract the correct lemma. Suffix rules are referred from Table 4.1, Table 4.2, Table 4.4 and Table 4.5 to generate a new word from nominalized/ derived word and to extract its lemma. (Francis Katamba. English Words, 1994).
    - Morphological Analysis of nominalized Noun/Adjective words.

i) Identification of Suffix and generation of token:

Lovins' stemming "Longest-match" principle is applied on input-word for suffix/endings identification (Noun Siffix: (e.g. -al, -ance, -ence, -ion, -ication, -ion, -ism, -ship, -(p)tion, -ure, -ment, -age, -cation, -ief; Adjective suffix: e.g. - able, -en, -ious, -ful, -AL, -ary, -ic, -ical, -y, -ed). The "longest-match principle" states that for any given class of endings (Noun class/Adjective class), if more than one ending provides a match, the one which has the longest match, should be selected and removed. The suffix "-ation" will get a higher priority to be removed as compared to suffix "-ion" (e.g. permutation).

If the length of identified suffix is greater than or equal to the length of truncated token, then further processing of selection of suffix is required before removal of suffix from input-word (e.g. National).

If the length of the truncated token is be greater than 4 and length of the selected suffix is less than the length of the token, then truncation is permitted to generate a stem-token.

ii) Extraction of Lemma:

If truncated stem-token exists as Verb POS word, then its lemma is extracted and selected as the lemma of input-word.

If the truncated stem-token does not exist as a valid dictionary word in any POS form, then the new token is generated using POS-class based recoding rule. Rules are shown in Table 4.1 to Table 4.8.

Recoding of token and construction of new word:

- Based on POS-class of the input-word and category of suffix, ending section of the stem-token is re-constructed to generate a new token.
- The new token is validated to check whether it is a valid dictionary word or not.
- If token is not identified as a dictionary word, process of re-construction of token is continued until valid dictionary word is generated.

  Else, input-word is identified as the lemma of the word.
- If new dictionary word exists either as a Verb or as a Noun form, the lemma of the word which is in Verb form is selected as the final lemma of the input-word.
- If new dictionary word exists either as a Verb or as an Adjective form, the lemma of the word which is in Verb form is selected as the final lemma of the input-word.
- If new dictionary word exists only in Noun form (e.g., historical → history, musical → music) and the length of the new dictionary Noun word is less than the length of the input-word, then the new word is selected as the final lemma.

Step: 4 –Parsing of word with Adverb POS tagged only.

If the input word only exists as an Adverb POS in any dictionary (e.g., gracefully, gently), it may exist as a nominalized adverb in English vocabulary.

Step: 4.1- Truncation of suffix and construction of a new word.

- If a new dictionary word is generated after applying Adverb suffix (e.g. -ly, -ally,-ably, -ically, -wise, -y) truncation rules and recoding rules, then, new word's lemma is selected as the lemma for input adverb word.
- Else, the input word (e.g., instead, fast) is selected as the lemma.

With the help of above mentioned steps and using the suffix and recoding rules, maximum allied nominalized words as well as allied morphed words are merged to their corresponding dictionary base words or morphological root words/lemma. With the help of this proposed model, construction of new dictionary root-wordis possible from any morphological derived input word. Most of the time, POS of input word and POS of the new word is different. Normal inflected nounwords (bird-birds/box-boxes) and irregular (fungus-fungi/woman-women) singular-plural noun, verb in any tense (run-ran/go-went) and all nominalized words (acceptance-acceptability/ application-applicability/ add-addition) can merge to their root words/lemmas using this proposed model.

Fig. 4.5 shows general morphological parsing process (FSA) of all input-words based on their POS class and suffix rules which are implemented in this LemmaChase model.

Table 4.1 shows suffix list of derived noun which are constructed from verb.

Table 4.1 Derivative Noun Suffix Rules from Verbs

Noun to Verb	**Input word**	**Lexical function within the item-and-process model**
-ation	don-ation, reconcili-ation, regul-ation, educ-ation	**[x]$_V$->[[x]$_V$*ation*]$_N$:** [donate]$_V$vation]$_N$,[regulate]$_V$vation]$_N$, educate]$_V$ion]$_N$
-al	Approv-al, arriv-al, revers-al, refus-al, remov-al	**[x]$_V$->[[x]$_V$*al*]$_N$:** **[[**Approve]$_V$val]$_N$,[arrive]$_V$val]$_N$,[reverse]$_V$al]$_N$
-ant	inhabit-ant, celebr-ant, assistant, attendant, lieutenant, consultant, accountant, contestant;	**[x]$_V$->[[x]$_V$*ant*]$_N$:**[inhabit]$_V$vant]$_N$, celebrate]$_V$vant]$_N$
-er	teach-er, runn-er, writ-er, build-er, paint-er	**[x]$_V$->[[x]$_V$*er*]$_N$,**:[[teach]$_V$*er*]$_N$,[[run]$_V$*er*]$_N$, [build]$_V$*er*]$_N$
-ist	cycl-ist, typ-ist, copy-ist	**[x]$_V$->[[x]$_V$*ist*]$_N$,**:[[cycle]$_V$ist]$_N$,[[type]$_V$ist]$_N$, [copy]$_V$ist]$_N$
-ion	eros-ion, persuas-ion (from persuade), radiat-ion,	**[x]$_V$->[[x]$_V$*ion*]$_N$,**: [[erode]$_V$ion]$_N$, [radiate]$_V$ion]$_N$,[[educate]$_V$ion]$_N$
-ment	pave-ment, appoint-ment, govern-ment, pay-ment	**[x]$_V$->[[x]$_V$*ment*]$_N$**: [appoint]$_V$*ment*]$_N$,
-ery	catt-ery, pigg-ery, bak-ery, cann-ery	**[x]$_V$->[[x]$_V$*ery*]$_N$:** [pig]$_V$very]$_N$
-ee	employ-ee, detain-ee, pay-ee, intern-ee	**[x]$_V$->[[x]$_V$*ee*]$_N$:** [employ]$_V$vee]$_N$
-ance	acceptance, assistance, resistance, admittance	**[x]$_V$->[[x]$_V$*ance*]$_N$:** [accept]$_V$vance]$_N$
-ence	*sentence, competence*; existence, insistence, persistence, coexistence; convergence, divergence, emergence, silence, valence, violence, science, patience, conscience, influence,	**[x]$_V$->[[x]$_V$*ee*]$_N$:** [exist]$_V$vence]$_N$ [x]$_N$-> [x]$_N$: [[~~sent]$_V$~~ence]$_N$

Table 4.2 shows suffix list of derived verb which are constructed from noun.

Table 4.2 Derivative Verb Suffix Rules from Noun

Verb to Noun	Input Word	Lexical function within the item-and-process model
-ate	regul-ate, capacit-ate, don-ate	**[x]$_{NOUN}$->[[x]$_{NOUN}$*ate*]$_V$:**[[regular]$_N$ate]$_V$, [[capacity]$_N$ate]$_V$
-ise/-ize	colon-ise, American-ise, computer-ise	**[x]$_{NOUN}$->[[x]$_{NOUN}$*ise*]$_V$:**[[computer]$_N$ize]$_V$
-ify	carbonify , historify, personify, solidify	**[x]$_{NOUN}$->[[x]$_{NOUN}$*ify*]$_V$:** [[carbon]$_N$ify]$_V$, [[history]$_N$ify]$_V$,
-ify	amplify, beautify, calcify, certify, clarify, classify, codify, crucify; dignify, dissatisfy, diversify, edify, electrify, emulsify, exemplify	**[x]$_{NOUN}$->[[x]$_{NOUN}$*ify*]$_V$:** [[certificate]$_N$ify]$_V$, [[electric$_N$ify]$_V$,
-er	bicker, chatter, clatter, clutter, falter, flatter, flicker, fluster, flutter, glimmer	**[x]$_{NOUN}$->[[x]$_{NOUN}$*er*]$_V$:** [[chat]$_N$er]$_V$

Table 4.3 shows suffix list of derived adjective which are generated from verb.

Table 4.3 Derivative Adjective Suffix Rules from Verb

Adjective To Verb	Input Word	Lexical function within the item-and-process model
-ing	wait-ing ,stand-ing	**[x]$_V$->[[x]$_V$*ing*]$_{ADJ}$:** [[wait]$_V$ing]$_{ADJ}$, [[stand]$_V$ing]$_{ADJ}$
-ise/-ize	real-ise, neutral-ise, fertil-ise, immun-ise	**[x]$_V$->[[x]$_V$*ise*]$_{ADJ}$:** [[neutral]$_N$ise]$_{ADJ}$,[[fertile]$_V$ise]$_{ADJ}$,
-ive	act-ive, pens-ive, indicat-ive, evas-ive, product-ive,	**[x]$_V$->[[x]$_V$*ive*]$_{ADJ}$:** [[act]$_V$ive]$_{ADJ}$, [[indicate]$_V$ise]$_{ADJ}$
-able	read-able, govern-able; manage-able, do-able	**[x]$_V$->[[x]$_V$*able*]$_{ADJ}$:** [read]$_V$able]$_{ADJ}$, [govern]$_V$able]$_{ADJ}$

Table 4.4 shows suffix list of derived adjective which are generated from adjective.

Table 4.4 Derivative Adjective Suffix Rules from Adjective

Adjective to Adjective	Input Word	Lexical function within the item-and-process model
-ish	narrow-ish, blu-ish, pink-ish	**[x]$_{ADJ}$->[[x]$_{ADJ}$*ish*]$_{ADJ}$:**[[narrow]$_{ADJ}$ish]$_{ADJ}$, [[blue]$_{ADJ}$ish]$_{ADJ}$

Table 4.5 shows suffix list of derived adjective which are constructed from noun.

Table 4.5 Derivative Adjective Suffix Rules from Noun

Adjective to Noun	Input Word	Lexical function within the item-and-process model
-al	autumn-al, dent-al, division-al, medicin-al, origin-al, univers-al	[x]$_{NOUN}$->[[x]$_{NOUN}$*al*]$_{ADJ}$: ~~[[dent]~~$_N$al]$_{ADJ}$, [[medicine]$_N$al]$_{ADJ}$, [[origin]$_N$al]$_{ADJ}$, [[universe]$_N$al]]$_{ADJ}$

-ate	intim-ate, accur-ate, obdur-ate,	[x]NOUN->[[x]NOUN*ate*]ADJ:[[intima*cy*]Nate]ADJ, [[accura*cy*]Nounte]ADJn,
-ish	lout-ish, fiend-ish, freak-ish, child-ish	[x]NOUN->[[x]NOUN*ish*]ADJ: [[child]Nish]ADJ, [[freak]Nish]ADJ
-less	joy-less, care-less, fear-less	[x]NOUN->[[x]NOUN*less*]ADJ: [[joy] NOUN less]ADJ
-ful	joy-ful, care-ful, fear-ful, cheer-ful	[[joy]Nful]ADJ, [[cheer]Nful]ADJ
-(i)an	Chomsky-an, Dominic-an, suburb-an,	[[reptile]Nian]ADJ, [[Canada]Nian]ADJ
-some	quarrel-some, trouble-some	[[quarrel]Nsome]ADJ, [trouble]Nsome]ADJ
-ic	fantastic, metallic, systemic	[[fantasy]Nic]ADJ,[[metal]Nic]ADJ, [[system]Nic]ADJ, [[allergy]Nic]ADJ
-ible	audible, credible, permissible, admissible	[x]NOUN->[[x]NOUN*ible*]ADJ: [[Audio]NOUN*ible*]ADJ
-ial	commercial, facial, colonial	[[x]NOUN*ial*]ADJ: [[Commerce]NOUN*ial*]ADJ
-an/-n	African, American	[[x]NOUN*an*]ADJ: [[Africa]NOUN*n*]ADJ
-ian/-n	Asian,	[[x]NOUN*ian*]ADJ: [[Asia]NOUN*n*]ADJ
-ant	Fragnant, elegant	[[x]NOUN*ant*]ADJ: [[Fragnance]NOUN*ant*]ADJ
-ary	Primary, fragmentary	[[x]NOUN*ary*]ADJ: [[Fragnance]NOUN*ary*]ADJ
-ory	Cursory, sensory	[[x]NOUN*ory*]ADJ: [[sensor]NOUN*ory*]ADJ
-ar	Circular, cellular, muscular	[[x]NOUN*ar*]ADJ: [[circle]NOUN*ar*]ADJ
-ical	aesthetical, allegorical, alphabetical , biological, botanical	[[x]NOUN*ical*]ADJ: [[alphabet]NOUN*icalt*]ADJ
istical	egoistical, egotistical, logistical, statistical.	[[x]NOUN*istical*]ADJ: [[statistics]NOUN*istical*]ADJ
-ly	costly, deadly, friendly, ghastly	[[x]NOUN*ily*]ADJ: [[cost]NOUN*ily*]ADJ
-ous	famous, jealous, zealous, luminous, mountainous, ominous, poisonous	[[x]NOUN*ous*]ADJ: [[mountain]NOUN*ous*]ADJ
-eous	erroneous, sebaceous, spontaneous; courageous	[[x]NOUNe*ous*]ADJ: [[error]NOUNe*ous*]ADJ
-y	angry, catchy, crafty, dreamy, empty	[[x]NOUNy]ADJ: [[dream]NOUNy]ADJ
-ern	eastern, northern, southern	[[x]NOUN*ern*]ADJ: [[east]NOUN*ern*]ADJ
-ine	alkaline, aniline, aquiline,	[[x]NOUN*ine*]ADJ: [[alkali]NOUN*ine*]ADJ
-ist	capitalist, chauvinist, communist, extremist	[[x]NOUN*ist*]ADJ: [[capital]NOUN*ist*]ADJ
-like	childlike, ghostlike, godlike	[[x]NOUN*like*]ADJ: [[child]NOUN*like*]ADJ
-some	awesome, burdensome, cumbersome	[[x]NOUN*some*]ADJ: [[]NOUN*some*]ADJ
-th	fourth, fifth, sixth, seventh, depth, width, length	[[x]NOUN*th*]ADJ: [[four]NOUN*th*]ADJ
-ward	backward, downward, forward	[[x]NOUN*ward*]ADJ: [[back]NOUN*ward*]ADJ

Table 4.6 shows suffix list of derived noun which are constructed from verb or noun.

Table 4.6 Derivative Noun Suffix Rules from Verb/Noun

Noun to Noun/Verb	**Input word**	**Lexical function within the item-and-process model**
-cy	privacy, legacy, fallacy, accuracy, adequacy,	[[x]NOUN*cy*]NOUN: [[accurate]NOUN*cy*]NOUN
-ade	brigade, grenade, parade, blockade	[[x]NOUN*ward*] NOUN: [[back]NOUN*ward*] NOUN
-age	blockage, bondage, breakage, carnage, courage, package, savage, damage, image	[[x]NOUN*age*] NOUN: [[back]NOUN*age*] NOUN, [[break]VERB*age*]NOUN
-al	hemical, festival, hospital, interval, terminal; approval, removal, referral, rehearsal, dismissal,	[[x]VERB*al*] NOUN:[[approve]VERB*al*] NOUN, [[remove]VERB*age*]NOUN
-n	African, Anglican, cardigan, talisman, curtain, fountain, mountain, German,	[[x]NOUN*n*] NOUN: [[Africa]NOUN*n*]NOUN [[xn]NOUN]NOUN:[Mountai~~n~~]Noun, [Germa~~n~~]Noun
-dom	freedom, kingdom, wisdom,	[[x]NOUN*dom*] NOUN: [[free]NOUN*dom*] NOUN, [[wise]NOUN*dom*]NOUN
-hood	childhood, babyhood, boyhood, girlhood	[[x]NOUN*hood*] NOUN: [[child]NOUN*hood*] NOUN, [[girl]NOUN*hood*]NOUN
-ic	arctic, classic, ethic, magic, music, rhetoric, arithmetic; characteristic	[[x]NOUN*ic*]NOUN: [[character]NOUN*ic*]NOUN
-ism	realism, communism, feudalism, nihilism, capitalism	[[x]NOUN*ism*]NOUN: [[back]NOUN*age*]NOUN, [[break]NOUN*age*]NOUN
-ist	dentist, typist, stylist, chemist, scientist, tourist	[[x]NOUN*ist*]NOUN: [[style]NOUN*ist*]NOUN,

		[[type]$_{NOUN}$*ist*]$_{NOUN}$
-ment	treatment; apartment, department, appointment, government; environment, assignment, alignment	[[x]$_{VERB}$*ment*]$_{NOUN}$: [[govern]$_{VERB}$*ment*]$_{NOUN}$, [[assign]$_{VERB}$*ment*]$_{NOUN}$
-ness	blindness, brightness, coldness, darkness, toughness	[[x]$_{NOUN}$*age*]$_{ness}$: [[blind]$_{NOUN}$*ness*]$_{NOUN}$, [[cold]$_{NOUN}$*ness*]$_{NOUN}$
-ship	friendship, hardship, worship, courtship, leadership	[[x]$_{NOUN}$*ship*]$_{NOUN}$: [[friend]$_{NOUN}$*ship*]$_{NOUN}$, [[leader]$_{NOUN}$*ship*]$_{NOUN}$
-th	breadth, depth, filth, growth, health, length	[[x]$_{NOUN}$*th*]$_{NOUN}$: [[deep]$_{NOUN}$*th*]$_{NOUN}$, [[grow]$_{VERB}$*th*]$_{NOUN}$
-ancy	poignancy, occupancy, militancy; discrepancy, expectancy, redundancy.	[[x]$_{VERB}$ *ancy*]$_{NOUN}$: [[occupy]$_{VERB}$*ancy*]$_{NOUN}$, [[expect]$_{VERB}$*ancy*]$_{NOUN}$
-cy	agency, urgency, tendency, clemency, currency	[[x]$_{NOUN}$*age*]$_{NOUN}$: [[agent]$_{NOUN}$*cy*]$_{NOUN}$, [[urgent]$_{NOUN}$*cy*]$_{NOUN}$
-dent	respondent, correspondent, superintendent; present, moment, patent, talent, tangent	[[x]$_{NOUN}$*dent*]$_{NOUN}$: [[response]$_{NOUN}$*dent*]$_{NOUN}$
-ary	secretary, dignitary, military, notary, votary, lapidary, dromedary, emissary, adversary	[[x]$_{NOUN}$*ary*]$_{NOUN}$: [[advice]$_{NOUN}$*ary*]$_{NOUN}$, [[dignity]$_{NOUN}$*ary*]$_{NOUN}$
-ery	archery, fishery, bravery, slavery, flattery, lottery, robbery, snobbery	[[x]$_{NOUN}$*ery*]$_{NOUN}$: [[fish]$_{NOUN}$*ery*]$_{NOUN}$, [[brave]$_{NOUN}$*ery*]$_{NOUN}$
-ry	memory, allegory, oratory, dormitory, lavatory	[[x]$_{NOUN}$*ry*]$_{NOUN}$: [[orator]$_{NOUN}$*ry*]$_{NOUN}$
-or	sailor, tailor, janitor, operator, aviator, navigator; contractor, director, inspector, inventor, investor	[[x]$_{VERB}$*or*]$_{NOUN}$: [[operate]$_{VERB}$*or*]$_{NOUN}$, [[inspect]$_{VERB}$*or*]$_{NOUN}$,
-ar	beggar, burglar, liar, scholar, vicar, cougar, dollar, calendar, circular	[[x]$_{VERB}$*ar*]$_{NOUN}$: [[beg]$_{VERB}$*ar*]$_{NOUN}$, [[lie]$_{VERB}$*ar*]$_{NOUN}$
-er	career, veneer, auctioneer, engineer, gazetteer, mountaineer,	[[x]$_{VERB}$*or*]$_{NOUN}$: [[engine]$_{NOUN}$*er*]$_{NOUN}$ [[inspect]$_{VERB}$*or*]$_{NOUN}$,
-ee	absentee, addressee, devotee, divorcee, employee, endorsee	[[x]$_{NOUN}$*ee*]$_{NOUN}$: [[absent]$_{NOUN}$*ee*]$_{NOUN}$ [[employ]$_{VERB}$*ee*]$_{NOUN}$,
-ess	actress, goddess, governess, hostess, mistress, poetess	[[x]$_{VERB}$*ess*]$_{NOUN}$: [[god]$_{NOUN}$*ess*]$_{NOUN}$ [[govern]$_{VERB}$*ess*]$_{NOUN}$,
-ion	completion, deletion, creation, equation, discretion, devotion, emotion, promotion, concoction; action, auction, faction, attraction, contraction, subtraction, distraction, extraction	[[x]$_{VERB}$*ion*]$_{NOUN}$: [[complete]$_{NOUN}$*ion*]$_{NOUN}$ [[promote]$_{VERB}$*ion*]$_{NOUN}$, [[substract]$_{VERB}$*ion*]$_{NOUN}$
-tion	addition, ambition, audition, condition, edition, rendition, tradition; erudition, expedition, extradition	[[x]$_{VERB}$*ion*]$_{NOUN}$: [[add]$_{NOUN}$*tion*]$_{NOUN}$ [[trade]$_{NOUN}$ *tion*]$_{NOUN}$, erude]$_{VERB}$ *tion*]$_{NOUN}$
	formation; information, reformation, transformation, confirmation, defamation, animation, automation,	[[x]$_{VERB}$*ation*]$_{NOUN}$: [[form]$_{VERB}$*ation*]$_{NOUN}$ [[confirm]$_{VERB}$*ation*]$_{NOUN}$, [[automate]$_{VERB}$*ation*]$_{NOUN}$
-ion	vision, lesion; revision, division, provision, collision, derision, decision, incision, precision, excision, adhesion, cohesion;	[[x]$_{VERB}$*ion*]$_{NOUN}$: [[decide]$_{NOUN}$*ion*]$_{NOUN}$ [[confuse]$_{VERB}$*ion*]$_{NOUN}$, [[excise]$_{VERB}$*ion*]$_{NOUN}$
-y	safety, nicety, deity, surety, liberty, poverty, property, puberty, honesty, majesty, certainty	[[x]$_{NOUN}$*or*]$_{NOUN}$: [[safe]$_{NOUN}$*ty*]$_{NOUN}$ [[honest] $_{NOUN}$*ty*]$_{NOUN}$, [[certain]$_{NOUN}$*ty*]$_{NOUN}$
-ibility	credibility, possibility, sensibility, flexibility, visibility, legibility; eligibility	[[x]$_{NOUN}$*ibility*]$_{NOUN}$: [[possible]$_{NOUN}$*ibility*]$_{NOUN}$ [[sense] $_{NOUN}$*ibility*]$_{NOUN}$, [[eligible]$_{NOUN}$*ibility*]$_{NOUN}$

Table 4.7 shows suffix list of derived verb which are constructed from verb.

Table 4.7 Derivative Verb Suffix Rules from Verb

Verb to Verb	Input Word	Lexical function within the item-and-process model
-er	chatt-er, patt-er, flutt-er	[[chat]$_V$er]$_V$, [[flut]$_V$er]$_V$
-ify	Sign-ify	[[sign]$_V$ify]$_V$

Table 4.8 shows suffix list of derived adverb which are constructed from adjective.

Table 4.8 Derivative Adverb Suffix Rules from Adjective

Adverb to Adjective	Input Word	Lexical function within the item-and-process model
-ly	usual-ly, busi-ly, proud-ly, loud-ly, grateful-ly, expensive-ly	**[x]$_{ADJ}$->[[x]$_{ADJ}$*ly*]$_{ADV}$:**[[busy]$_{ADJ}$ly]$_{ADV}$,[[loud]$_{ADJ}$ly]$_N$, [[expensive]$_{ADJ}$ ly]$_{ADV}$
-ly	brutally, centrally, dismally, equally, fatally, finally, formally, frugally, globally	**[x]$_{ADJ}$->[[x]$_{ADJ}$*ly*]$_{ADV}$:**[[equal]$_{ADJ}$ly]$_{ADV}$,[[central]$_{ADJ}$ly]$_{ADV}$
-ally	additionally, conditionally, conventionally, emotionally, exceptionally, intentionally	**[x]$_{ADJ}$->[[x]$_{ADJ}$*ally*]$_{ADV}$:**[[addition]$_{ADJ}$ally]$_{ADV}$,[[exception]$_{ADJ}$ally]$_{ADV}$
-ly	cordially, jovially, partially, trivially; crucially, racially, socially, specially; commercially	**[x]$_{ADJ}$->[[x]$_{ADJ}$*ly*]$_{ADV}$:**[[social]$_{ADJ}$ly]$_{ADV}$,[[commercial]$_{ADJ}$ally]$_{ADV}$
-ly	accurately, adequately, delicately, intricately, desperately, moderately, separately; intimately, ultimately, fortunately	**[x]$_{ADJ}$->[[x]$_{ADJ}$*ly*]$_{ADV}$:**[[delicate]$_{ADJ}$ly]$_{ADV}$,[[moderate]$_{ADJ}$ally]$_{ADV}$
-fully	awfully, carefully, cheerfully, doubtfully, dreadfully, faithfully, fearfully	**[x]$_{ADJ}$->[[x]$_{ADJ}$*fully*]$_{ADV}$:** [[cheer]$_{ADJ}$fully]$_{ADV}$,[[care]$_{ADJ}$fully]$_{ADV}$

Table 4.9 shows suffix list of derived noun which are constructed from noun.

Table 4.9 Derivative Noun Suffix Rules from Noun

Noun to Noun	Input Word	Lexical function within the item-and-process model
-aire	million-aire, doctrin-aire, solit-aire	**[[x]$_{NOUN}$*aire*]$_{NOUN}$** : [[million]$_N$ aire]$_N$
-acy	advoc-acy, episcop-acy, intim-acy, accur-acy,	**[[x]$_{NOUN}$*acy*]$_{NOUN}$** : [[advocate]$_N$ acy]$_N$, [[accurate]$_N$ acy]$_N$
-er	marin-er, geograph-er, football-er	**[[x]$_{NOUN}$*er*]$_{NOUN}$** : [[football]$_N$ er]$_N$, [[geograph]$_N$ er]$_N$
-ery	machin-ery, crock-ery, jewell-ery, pott-ery	**[[x]$_{NOUN}$*ery*]$_{NOUN}$**: [[machine]$_N$ ery]$_N$, [[crock]$_N$ ery]$_N$, [[jewel]$_N$ ery]$_N$
-let	pig-let, is-let, riv(u)-let	**[[x]$_{NOUN}$*let*]$_{NOUN}$** : [[pig]$_N$ let]$_N$
-ling	duck-ling, prince-ling, found-ling	**[[x]$_{NOUN}$*ling*]$_{NOUN}$**: [[duck]$_N$ ling]$_N$, [[prince]$_N$ ling]$_N$
-ship	king-ship, craftsman-ship, director-ship	**[[x]$_{NOUN}$*ship*]$_{NOUN}$** : [[king]$_N$ship]$_N$, [[director]$_N$ ship]$_N$
-ism	femin-ism, capital-ism, Marx-ism, structural-ism	**[[x]$_{NOUN}$*ism*]$_{NOUN}$** : [[feminine]$_N$ism]$_N$, [[capital]$_N$ism]$_N$,**[[[structure]$_N$ al]$_{ADJ}$]ism]$_N$**
-ist	femin-ist, capital-ist, Marx-ist, structural-ist	**[[x]$_{NOUN}$*ist*]$_{NOUN}$** [[feminine]$_N$ist]$_N$,[[[structure]$_N$ al]$_{ADJ}$]ist]$_N$

4.5 Parsing of Surface Word

"Finite-State Morphological Analysis" approach is one of the best processes to parse an input word. Morphology is the area of linguistics that explores the construction of words. It distinguishes between surface word and their analyses, called lemmas. The lemma for a surface form such as the English word "developmental" typically consists of the traditional dictionary form of the word together with terms that convey the morphological properties of the particular form. The lemma for "developmental" might be represented as "develop$_{Verb}$+ment$_{Noun}$+al$_{Adjective}$" to indicate that "developmental" is the derivative form of the verb "develop".

At the time of designing of this LemmaChase model, finite-state approach is followed to generate the lemma from any derived English word. The basic claim of the finite-state approach to morphology is that the relation between the surface word of a language and their corresponding lemmas can be described as a regular relation[7].

Finite State Transducer (FSAutomaton) technique is used in finite-state morphological approach. That transducer technique is partially incorporated to develop this LemmaChase model. Finite-state transducers are often used for morphological analysis in NLP and linguistic research and applications.

FSAutomaton concept:

• FSAutomaton is used to recognize the string which is accepted as an input. The automaton computes a function that maps strings into the set {0, 1}. An automaton generates strings, which will be an output tape.

• An FSA represents a set of strings. e.g. {develop, develops, developed, development, developmental}

• A recognizer function. Function recognize (str) return true or false.

• FSTransducers : An FST represents a set of pairs of strings (input, output pairs). e.g.

{(develop, develop+V+ed), (develop+N+ment), (develop+N+ment+ADJ+al) ...}

• A transducer function: It maps input to zero or more outputs.

transduce(develop) -> { develop+V+ed, develop+N+ment} can return multiple answers .

- FSAutomata have input labels.
- FSTransducers have input: output pairs on labels.

Below mentioned symbols are used for FS Transducer.

[7] https://web.stanford.edu/~laurik/publications/ciaa-2000/fst-in-nlp/fst-in-nlp.html

Q: a finite set of N states $q_0, q_1, \ldots, q_{N-1}$
Σ: a finite set corresponding to the input alphabet

$q_1 \in Q$: the start-state
Γ : finite set, called the output alphabet;
I : subset of Q, the set of initial states;

$F \subseteq Q$ the set of final states;
$\delta(q,w)$ the transition function or transition matrix between states;

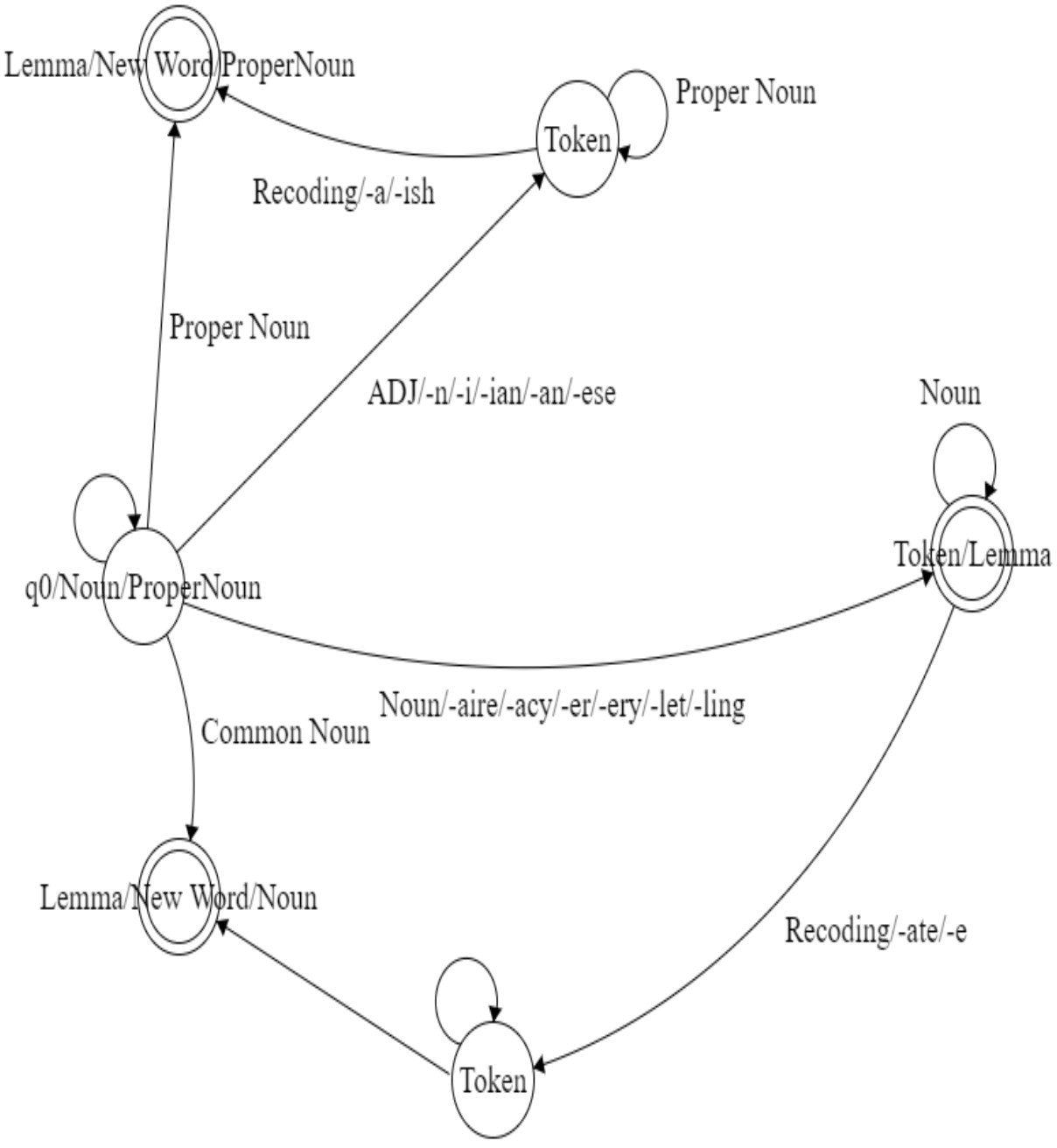
Given a state $q \in Q$ and a string $w \in \Sigma^*$, $\delta(q,w)$ returns a set of new states

The word-parsing state transition diagram given in Fig4.5 follows the above mentioned set of rules.

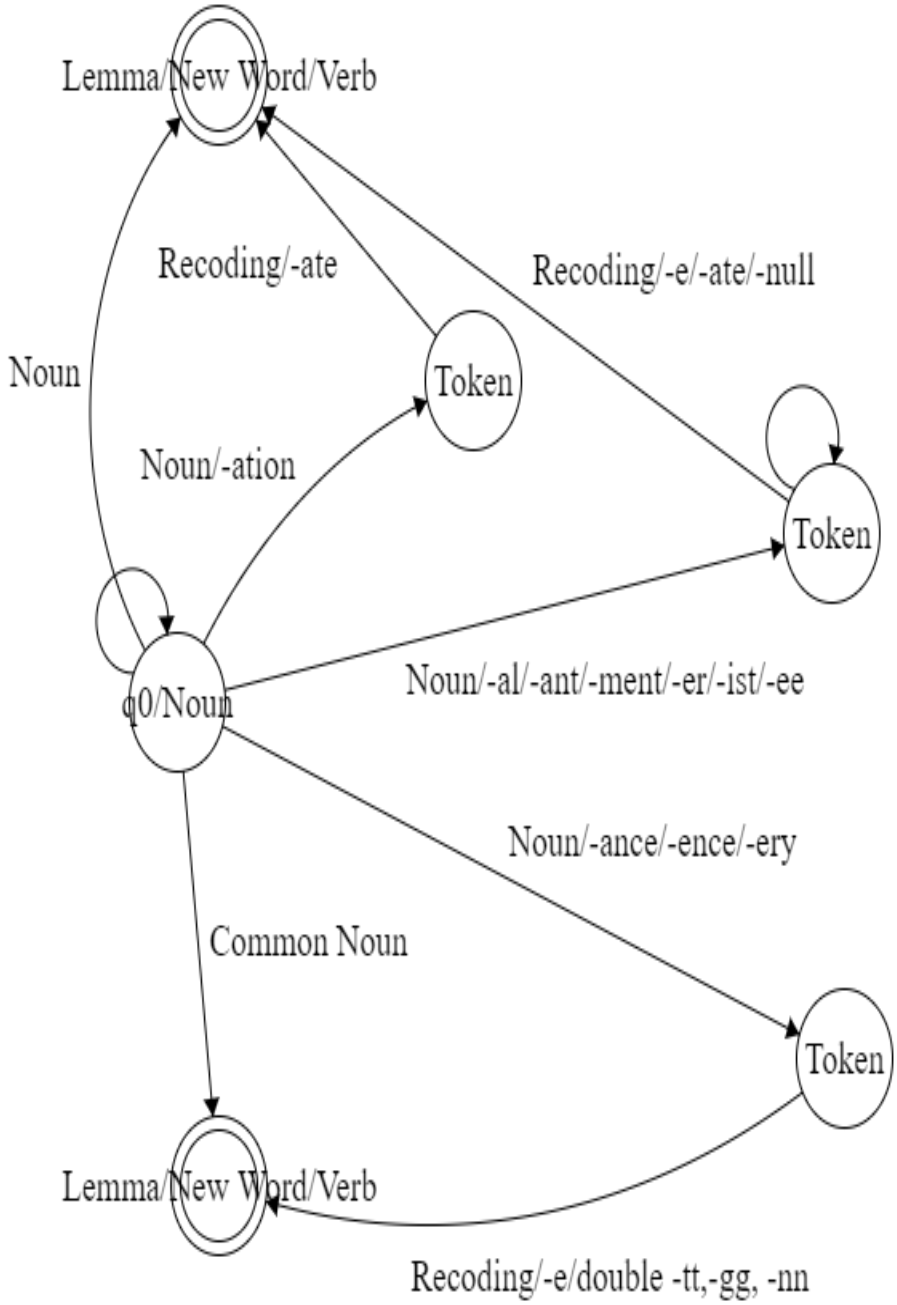
- Starting state indicates q_0 state, surface input word's state.
- The transition states are represented as "Token-POS"'s state (q_i).
- $\delta(q,w)$ returns a set of states of newly generated token with new POS .
- $\delta(q, w)$ computes suffix truncation and recoding process for surface input word to generate a new **Lemma** of a definite POS, different from input word's POS.

Fig 4.5 shows word state transition diagram through which any word will move towards the root word.

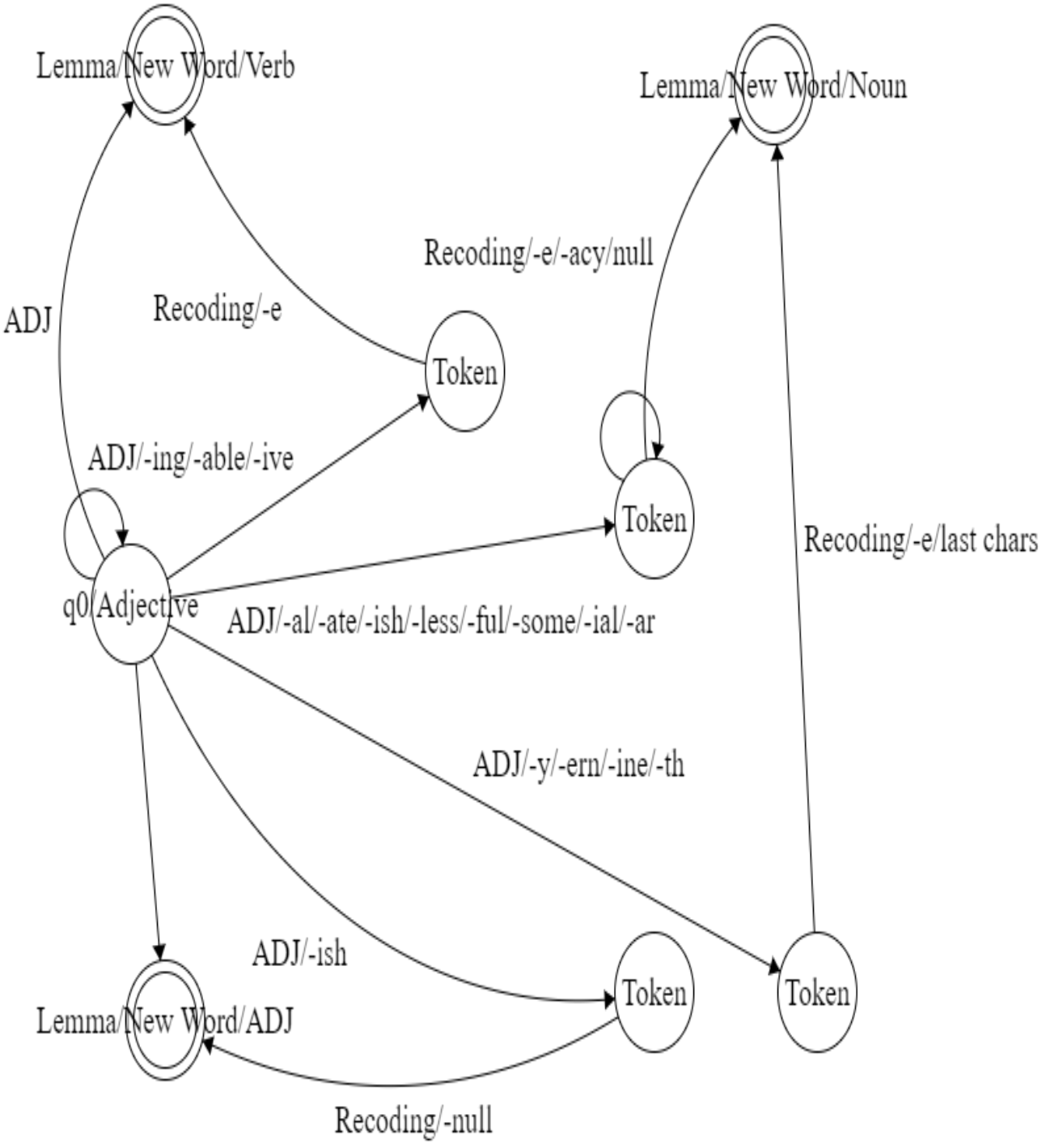
State Transition from Derivative Noun to Noun lemma and Derivative Adjective to Proper Noun: Nominalized Word Parsing



State Transition from Derivative Noun to Verb Lemma: Nominalized Word Parsing



State Transition of Derivative Adjective Word to Verb/Noun/Adjective Lemma : Nominalized Word Parsing



State Transition of Adverb Derivative Word to Adjective Lemma: Nominalized Word Parsing

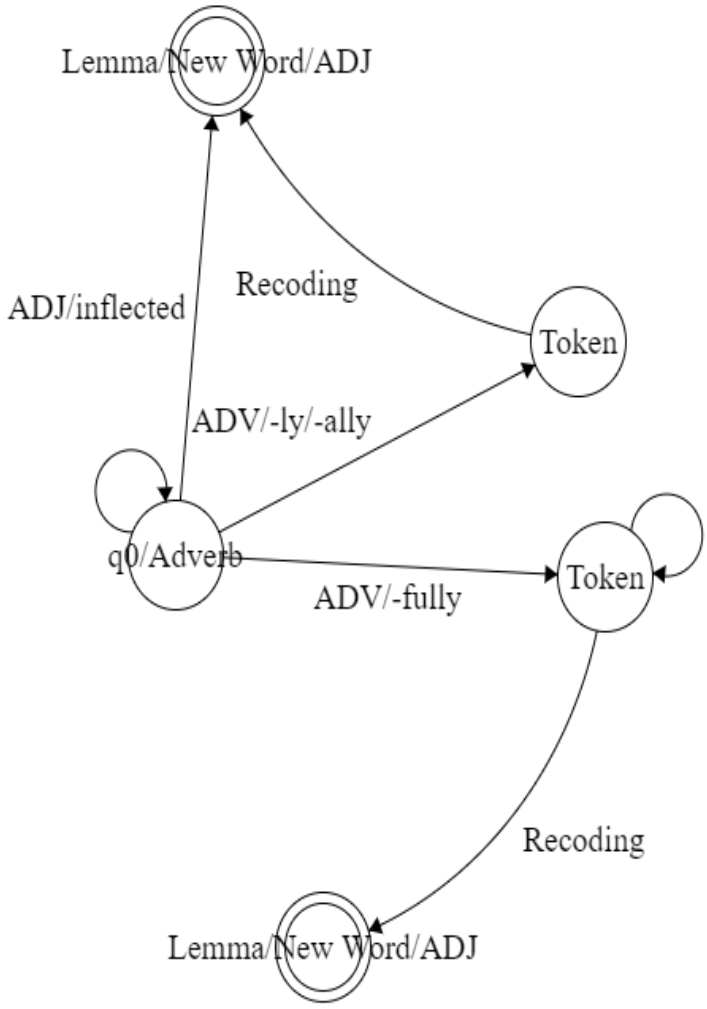


Figure 4.5 State Transition of Input Words

The algorithm of LemmaChase explains the flow of the transition of an input word to reach to the final state of its lemma form through processing of different stem token.

4.6 Tools used in LemmaChase

POS-tagger (Christopher D. Manning: Stanford), JWNL tool (Computing with a Thesaurus Word Senses and Word Relations)[8] and WordNet dictionary (Jorge Morato, Miguel Ángel Marzal, Juan Lloréns, and José Moreiro) are used to identify the context and to

[8] https://web.stanford.edu/~jurafsky/slp3/slides/Chapter18_introandsimilarity.pdf

extract the lemma of a word. These tools assist theLemmaChase model in generating the correct lemma for any input word.

Part-Of-Speech Tagger: Stanford Lemmatizer uses Penn Treebank for tagging POS of each input-word of the proposed model. The Stanford POS-tagger is used in both proposed models "LemmaChase" and "LemmaQuest" (as discussed in Chapter 5) lemmatizers for identifying POS of each word.

Table 4.10 depicts sample input text.

Table 4.10 Sample Input Text for POS Tagging

"Eight Dead, Up to 18 Missing After Explosion at Marine Band ComplexEds: LEADS with 7 grafs to UPDATE with eight dead, Scotland Yard sending anti-terrorist unit to investigate; pickup 7th graf pvs, `Ten doctors."

Table 4.11 depicts sample output which was generated by Stanford POS-Tagger.

Table 4.11 Output of Stanford POS-Tagger

"Eight/CD Dead/JJ ,/, Up/IN to/TO 18/CD Missing/VBG After/IN Explosion/NN at/IN Marine/NNP Band/NNP ComplexEds/NNPS :/: LEADS/VBZ with/IN 7/CD grafs/NNS to/TO UPDATE/VB with/IN eight/CD dead/JJ ,/, Scotland/NNP Yard/NNP sending/VBG anti-terrorist/JJ unit/NN to/TO investigate/VB ;/: pickup/NN 7th/JJ graf/NN pvs/NNS ,/, `/`` Ten/CD doctors/NNS .../: '/"

Penn Treebank: The Penn Treebank, in its eight years of operation (1989-1996), produced approximately 7 million words of part-of-speech tagged text, 3 million words of skeletally parsed text, over 2 million words of text parsed for predicate argument structure, and 1.6 million words of transcribed spoken text annotated for speech disfluencies. All available Penn Treebank materials are distributed by the Linguistic Data Consortium[9].

Table 4.12 shows Penn bank POS tag list.

Table 4.12 Sample of POS Tag Set

POS tag	POS tag	POS tag
1. CC Coordinating conjunction	10.LS List item marker	19.PP Possessive pronoun
2. CD Cardinal number	11.MD Modal	20.RB Adverb
3. DT Determiner	12.NN Noun, singular or mass	21.RBR Adverb, comparative
4. EX Existential there	13.NNS Noun, plural	22.RBS Adverb, superlative
5. FW Foreign word	14.NNP Proper noun, singular	23.RP Particle
6. IN Preposition/subord.	15.NNPS Proper noun, plural	24.VBG Verb, gerund/present participle
7. JJ Adjective	16.PDT Predeterminer	25.VBP Verb, non-3rd ps. sing. present
8. JJR Adjective, comparative	17.POS Possessive ending	26.VBD Verb, past tense
9. JJS Adjective, superlative	18.PRP Personal pronoun	27.VB Verb, base form

[9] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

JWNL Tool: JWNL is an API for accessing WordNet-style relational dictionaries. It also provides functionality for relationship discovery and morphological processing. First, JWNL.initialize() is invoked to initialize the code of a program. Then, Dictionary.getInstance() is called to get the currently installed dictionary.

WordNet: WordNet was developed by Princeton University. WordNet 2.1 is used in this research work. "WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. Structure: Synonyms words that denote the same concept and are interchangeable in many contexts are grouped into unordered sets (synsets). Each of WordNet's 117000 synsets is linked to other synsets by means of a small number of "conceptual relations." [10]

4.7 Result and Discussion

This section discusses the output generated by LemmaChase and the output is compared with that of existing popular lemmatizers and morphological analyzers.

Output of LemmaChase: For derived and nominalized words in Noun, Adjective and Adverb form, LemmaChase generates output with 5% to 12% average error (Rupam Gupta, and Anjali G. Jivani. LemmaChase, 2020). Lemma generated by LemmaChase is shown in Table 4.13 for inflected and derived dictionary word set. All incorrect lemmas are shown in red in the tables given below.

Table 4.13 Sample Data Set of Derived and Inflected Words (Mixed POS) with Lemmas

Seq. No	Input Word	Lemma	Input Word	Lemma	Input Word	Lemma
1.	abactor	abactor	Abattoirs	abattoir	Abdominal	Abdomen
2.	abactors	abactor	Abattoir	abattoir	Abdomen	Abdomen
3.	abacuses	abacus	Abbacies	abbacy	Abdomens	Abdomen
4.	abacus	abacus	Abbacy	abate	Abdominally	Abdominal
5.	abandonment	abandon	Abbeys	abbey	Abdominal	Abdominal
6.	abandons	abandon	Abbey	abbey	Abductor	Abduct
7.	abandon	abandon	Abbesses	abbess	Abducts	Abduct
8.	abandoned	abandon	Abbess	abbess	Abduction	Abduct
9.	abandoning	abandon	Abbots	abbot	Abducted	Abduct
10.	abandonee	abandonee	Abbot	abbot	Abducting	Abduct
11.	abandonees	abandonee	Abbotsbury	abbotsbury	Abductors	Abduct
12.	abasement	abase	Abbott	abbott	Abductions	Abduct
13.	abased	abase	Abbreviating	abbreviate	Abduct	Abduct
14.	abasing	abase	Abbreviation	abbreviate	Abecedarians	Abecedarian
15.	abases	abase	Abbreviator	abbreviate	Abecedarian	Abecedarian
16.	abase	abase	Abbreviators	abbreviate	Abolishes	Abolish
17.	abashing	abash	Abbreviates	abbreviate	Abolishing	Abolish

[10] https://wordnet.princeton.edu/

18.	abashed	abash	Abbreviated	abbreviate	Abolished	Abolish
19.	abashes	abash	Abbreviate	abbreviate	Abolish	Abolish
20.	abash	abash	abbreviations	abbreviate	Abolishment	Abolish
21.	abated	abate	abdicant	abdicant	Abolitionists	Abolition
22.	abates	abate	abdicants	abdicants	Abolitionist	Abolition
23.	abators	abate	abdicators	abdicate	Abolitionism	Abolition
24.	Abatement	abate	abdicate	abdicate	Abolition	Abolition
25.	Abating	abate	abdicates	abdicate	Abominably	Abominably
26.	abate	abate	abdicated	abdicate	Abominate	Abominate
27.	abatements	abate	abdication	abdicate	Abominating	Abominate
28.	abator	abate	abdicating	abdicate	Abominated	Abominate
29.	abater	abater	abdications	abdicate	Abomination	Abominate
30.	abaters	abaters	abdicator	abdicate	Abominates	abominate
31.	abominations	abominate	absolute	absolute	Abstain	abstain
32.	abominators	abominate	absolutely	absolute	Abstainers	abstain
33.	abominator	abominate	absolutism	absolute	Abstention	abstention
34.	abominable	abominate	absolutist	absolute	Abstentions	abstention
35.	aboriginal	aborigine	absolved	absolve	Abstinence	abstin
36.	aborigine	aborigine	absolves	absolve	Abstinent	abstinent
37.	aborigines	aborigine	absolver	absolve	Abstinently	abstinent
38.	aborts	abort	absolving	absolve	Abstractive	abstract
39.	abortion	abort	absolve	absolve	Abstracter	abstract
40.	abort	abort	absolvers	absolve	Abstracted	abstract
41.	aborting	abort	absorbable	absorb	Abstractor	abstract
42.	aborted	abort	absorbed	absorb	Abstracts	abstract
43.	abortions	abort	absorbs	absorb	Abstractly	abstract
44.	abortive	abort	absorb	absorb	Abstracting	abstract
45.	abortively	abortive abort	absorption	absorption	Abstraction	abstract
46.	abortus	abortus	absorptive	absorptive	Abstractionist	abstract
47.	abounds	abound	abstaining	abstain	Abstractionism	abstract
48.	abound	abound	abstained	abstain	Abstractionists	abstract
49.	abounding	abound	abstains	abstain	Abstractions	abstract
50.	abounded	abound	abstainer	abstain	Abstractedness	abstract
51.	absents	absent	abstractors	abstract	Abstain	abstain
52.	absinth	absinth	abstracters	abstract	Abstainers	abstain
53.	absinths	absinth	abstract	abstract	Abstention	abstention
54.	absinthes	absinthe	abstractedly	abstract	Abstentions	abstention
55.	absinthe	absinthe	abstractness	abstract	Abstinence	abstin
56.	absoluteness	absolute	abstruseness	abstruse	Abstinent	abstinent
57.	absolutions	absolute	abstrusely	abstruse	Abstinently	abstinent
58.	absolutes	absolute	abstruse	abstruse	Abstractive	abstract
59.	absolutists	absolute	absurd	absurd	Abstracter	abstract
60.	absolution	absolute	absurdly	absurd	Abstracted	abstract
61.	abundance	abund	absurdity	absurd	Abusively	abusive abuse
62.	abundantly	abundant	absurdities	absurdity absurd	Abusiveness	abusive abuse
63.	abundant	abundant	absurdum	absurdum	Academy	academy
64.	abusing	abuse	abundance	abund	Academia	academia
65.	abuse	abuse	abundantly	abundant	Academician	academic academy
66.	abuses	abuse	abundant	abundant	Academicians	academic academy
67.	abuser	abuse	abusing	abuse	Academically	academic academy
68.	abused	abuse	abuse	abuse	Academies	Academy
69.	abusive	abuse	abuses	abuse	Academic	Academy
70.	abusers	abuse	abuser	abuse	Academics	Academy
71.	abusively	abusive abuse	abused	abuse	Accurate	Accurate
72.	abundance	abund	abusive	abuse	Accelerating	Accelerate
73.	abundantly	abundant	abusers	abuse	Acclimatizing	Acclimatize
74.	accelerometers	accelerometer	accessibility	accessibility	Acclimatize	Acclimatize
75.	accents	accent	accessorise	accessorise	Acclimatizes	Acclimatize
76.	accented	accent	accessorize	accessorize	Acclimatized	Acclimatize
77.	accenting	accent	accessories	accessory access	Acclimatization	Acclimatize
78.	accent	accent	accidence	accidence	Accommodative	Accommodate
79.	accentual	accentual	accidens	accidens	Accommodate	Accommodate
80.	acceptability	accept	accidentals	accident	Accommodating	Accommodate
81.	accepting	accept	accidentally	accident	Accommodations	accommodate
82.	acceptable	accept	accident	accident	Accommodation	accommodate
83.	acceptance	accept	accidental	accident	Accommodatingly	accommodate
84.	acceptor	accept	accidents	accident	Accommodated	accommodate
85.	acceptors	accept	acclaim	acclaim	Accommodates	accommodate
86.	acceptances	accept	acclaims	acclaim	Accompanies	accompany
87.	accepts	accept	acclaiming	acclaim	Accompanied	accompany
88.	acceptant	accept	acclaimed	acclaim	Accompaniments	accompany

89.	accept	accept	acclamation	acclamation	Accompaniment	accompany
90.	acceptation	accept	acclamations	acclamation	Accomplices	accomplice
91.	accepted	accept	acclamatory	acclamatory	Accomplice	accomplice
92.	acceptably	acceptably	acclimate	acclimate	Accomplishes	accomplish
93.	accessibly	access	acclimating	acclimate	Accomplished	accomplish
94.	accessible	access	acclimation	acclimate	Accomplish	accomplish
95.	accessed	access	acclimates	acclimate	Accomplishable	accomplish
96.	accesses	access	acclimated	acclimate	Accomplishment	accomplish
97.	accession	access	acclimatised	acclimatise	Accomplishments	accomplish
98.	accessing	access	acclimatises	acclimatise	Accomplishing	accomplish
99.	access	access	acclimatise	acclimatise	Accords	accord
100.	accessory	access	acclimatising	acclimatise	Accordant	accord
101.	accessions	access	acclimatisation	acclimatise	Accord	accord
102.	acquaintances	acquaint	acquaintanceship	acquaint	Acquainting	acquaint
103.	acquaintance	acquaint	acquainted	acquaint	Acquaintanceship	acquaint

Observation:
Number of Derived and Inflected morphed words : 418 ; Total number of lemma generated: 123
Incorrect lemma generated : 25 errors Out of 418 words; Error rate :5.9%

Table 4.14 depicts the sample output with a set of derived Noun words input.

Table 4.14 Sample Data Set of Derived Noun Words with Lemmas

	Input-Word	Lemma	Input-Word	Lemma	Input-Word	Lemma	Input-Word	Lemma
1.	abrasive	abrase	courtship	court	immediacy	immediate	Packet	packet
2.	acceptance	accept	courtesan	courtesan	importance	import	parliamentarian	parliamentarian
3.	accuracy	accurate	credentials	credential	Incompetence	incompet	partisan	partisan
4.	acoustics	acoustic	criminal	criminal	indulgence	indulge	partnership	partner partner
5.	acrobatics	acrobatics	damage	damage	inheritance	inherit	pediatrics	pediatrics
6.	adage	adage	darkness	dark	insistence	insist	pellet	pellet
7.	additive	add	dealership	dealer	installment	install	persistence	persist
8.	adequacy	adequate	defensive	defense	instance	instance	phonetics	phonetics
9.	adhesive	adhesive	delicacy	delicate	intellectual	intellectual	physique	physique
10.	adjustment	adjust	dentist	dent	intelligence	intelligence	development	develop
11.	admittance	admit	department	department	intimacy	intimacy	sentence	sentence
12.	alignment	align	depth	depth	intricacy;	intricate	imitation	imitate
13.	alley	alley	detective	detect	intricacy	intricate	pocket	pocket
14.	analyst	analyst	dictatorship	dictator dictate	investment	invest	politics	politics
15.	analytics	analytics	diligence	diligence	italics	italy	practice	practice
16.	animal	animal	diplomacy	diplomacy	journalist	journal	privacy	private
17.	annulment	annul	dismissal	dismiss	journey	journey	procedure	procedure
18.	antarctic	antarctic	disposal	dispose	justice	justice	professional	profession profess
19.	apartment	apart	distance	distance	kidney	kidney	proposal	propose
20.	appointment	appoint	divergence	diverge	kindness	kind	public	public
21.	appraisal	appraise	doctrine	doctrine	leadership	leader	pudding	pudding
22.	approval	approve	dressing	dress	legacy	legate	quickness	quick
23.	arctic	arctic	dwelling	dwell	leisure	leis	rationalism	rational ration
24.	arithmetic	arithmetic	economics	economics	librarian	library	realism	real
25.	arrival	arrive	economist	economy	linguistics	linguistics	referral	referred refer
26.	artisan	artisan	emergence	emerge	literacy	literate	refusal	refuse
27.	assignment	assign	engine	engine	length	long	rehearsal	rehears
28.	assistance	assist	enlistment	enlist	longing	long	relationship	relation relate
29.	athletics	athletics	entertainment	entertain	magic	magic	reluctance	reluct
30.	babyhood	babyhood	environment	environment	mallet	mallet	remittance	remit
31.	bandage	bandage	ethic	ethic	marshal	marshal	removal	remove
32.	barbarian	Barbary	ethics	ethic	materialism	material matter	representative	representative
33.	basket	basket	evening	even	mathematics	mathematics	republic	republic
34.	blanket	blanket	executive	execute	meaning	mean	republican	republic
35.	blindness	blind	existence	exist	measure	measure	resistance	resist
36.	botanist	botany	explosive	explosive	mechanics	mechanic	revival	revive
37.	boyhood	boyhood	failure	fail	medal	medal	rhetoric	rhetoric
38.	breadth	bread	fallacy	fallacy	medicine	medicine	sardine	sardine
39.	brightness	bright	famine	famine	membership	member	sausage	sausage

40.	brotherhood	brotherhood	Feeling	feel	Metal	metal	savage	savage
41.	Building	build	feudalism	feudal	Midget	midget	Scholarship	scholar
42.	Bullet	bullet	Figure	figure	Money	money	Scientist	scientist
43.	Cabinet	cabinet	Filling	fill	Monopolist	monopoly	seizure	seize
44.	calisthenics	calisthenics	Filth	filth	Month	month	sentence	sentence
45.	Canine	canine	Friendship	friend	morning	morning	Service	service
46.	capitalism	capital	fulfillment	fulfill fulfill	Music	music	Shilling	shill
47.	Captive	captive	Galley	galle	narrative	narrate narrate	Sickness	sick
48.	cardinal	cardinal	garbage	garb	nationalism	national nation	Socialism	social
49.	Catalyst	catalyst	general	general	negligence	negligence	statesmanship	statesman
50.	Ceiling	ceiling	girlhood	girlhood	Nihilism	nihil	Statistics	statistic
51.	championship	champion champ	Goodness	good	Notice	notice	Stealth	steal
52.	characteristic	characteristic	government	govern	Novelist	novel	Stiffness	stiff
53.	Chemist	chemist	grammarian	grammar	Nugget	nugget	Strength	strength
54.	Childhood	childhood	Growth	grow	obstinacy	obstinate	Stylist	style
55.	Chimney	chimne	Guidance	guide	Offensive	offense offense	Subsistence	subsist subsist
56.	Citizenship	citizen	gymnastics	gymnastics	Office	office	Substance	subst
57.	Classic	class	Hardship	hard	Owlet	owlet	supremacy	supremate
58.	Closure	closure	Harshness	harsh	Ownership	owner	Survival	survive survive
59.	coexistence	coexist	Hatchet	hatchet	package	package	tenure	tenure
60.	Coldness	cold	Health	heal	Packet	packet	Thickness	thick
61.	Collateral	collateral	Heroine	heroine	Parliamentarian	parliamentarian	Toughness	tough
62.	Comet	come	Highness	high	partisan	partisan	Tourist	tour
63.	commitment	commit	Historian	history	Partnership	partner	treasure	treasure
64.	communism	commune	Hockey	hockey	Pediatrics	pediatrics	treatment	treat
65.	companionship	companion	Honey	honey	Pellet	pellet	Truth	truth
66.	Competence	compete	hostage	host	Persistence	persist	Typist	typist
67.	concealment	conceal	humanitarian	human	phonetics	phonetics	Urine	urine
68.	confederacy	confederate	hysterics	hysteric	physics	physic	vegetarian	vegetarian
69.	conspiracy	conspiracy	Idealism	ideal	Picket	picket	visage	visage
70.	convergence	converge	illiteracy	illiterate	Planet	plane	Voyage	voyage
71.	Copyist	copy	Illness	ill	pleasure	please	Warmth	warm
72.	courage	courage	image	image	Pocket	pocket	Wealth	wealth

Observation:
Number of Derived Noun words : 278 **; Total number of lemma generated:** 274
Error rate :12%

Table 4.15 shows derived adjective word set with their lemmas.

Table 4.15 Sample Data Set of Derived Adjective Words with Lemma

	Input-Word	**Lemma**	**Input-Word**	**Lemma**	**Input-Word**	**Lemma**
1.	academic	academy	bridal	bride	foolish	fool
2.	accessible	access	burly	bur	forceful	force
3.	acoustic	acoustic	capable	cap	formidable	form
4.	acrobatic	acrobat acrobat	Caspian	caspian	fourth	four
5.	admirable	admire	charitable	charity	friendly	friend
6.	aesthetic	aesthete	charming	charm	garish	gari
7.	aesthetical	aesthete	Christian	christ	Georgian	georgia
8.	affectionate	affection affect	collective	collect	German	german
9.	african	african africa	comely	come	ghastly	ghastly
10.	agnostic	agnostic	comfortable	comfort	ghostly	ghost
11.	alcoholic	alcohol	commercial	commerce	godly	god
12.	alkaline	alkaline	comparable	compare	graceful	grace
13.	allegorical	allegory	compatible	compatic	graduate	graduate
14.	allergic	allergy	convertible	convert	grateful	grate
15.	alphabetical	alphabet	costly	cost	homely	home
16.	amiable	amiable	countable	county	horrible	horrible

17.	amicable	amicable	creative	create	hospitable	hospitable
18.	analogical	analogy	credible	cred	humanistic	humanist human
19.	anatomical	Anatomical	creditable	credit	idealistic	idealist ideal
20.	anglican	anglican	cunning	cunning	imperialistic	imperial
21.	aniline	aniline	curable	cure	impressionistic	impress
22.	applicable	apply	deadly	dead	Indian	indian
23.	appropriate	appropriate	deceptive	deceptive	individualistic	individual
24.	approximate	approximate	defective	defect	innate	inn
25.	aquiline	aquiline	detective	detect	intelligible	intelligible
26.	archaeological	archaeology	directive	direct	irate	Ira
27.	arguable	argue	eastern	eastern	irritable	irritate
28.	articulate	articulate	eatable	eatable	journalistic	journalist journal
29.	asian	asian	edible	edible	judicial	judicial
30.	associate	associate	effective	effect	kindly	kind
31.	astronomical	astronomy	eighth	eighth	laughable	laugh
32.	attentive	attend	eleventh	eleven	lavish	lavish
33.	audible	audible	eligible	eligible	legible	legible
34.	lively	live	palpable	pay	liable	liable
35.	logistic	logistic	peevish	peeve	likely	like
36.	lonely	lone	perceptive	percept	likable	like
37.	lovable	love	Persian	persia	linguistic	linguist
38.	martian	mart	plausible	play	moralistic	moralist moral
39.	materialistic	material matter	portable	port	movable	movable
40.	mexican	mexican	possible	possible	ninth	ninth
41.	modern	Modern	Modern	modern	modern	modern
42.	moralistic	Moralistic	Moralistic	moralistic	moralistic	moralistic
43.	russian	Russian	Russian	russian	russian	russian
Observation: **Total number of Adjective words : 175, Number of lemma generated: 173** **Rate of error : 15%**						

Table 4.16 shows the list of mixed POS words with their Lemmas.

Table 4.16 Sample Data Set of Mixed POS Words with Lemmas

	Input-word	Lemma	Input-Word	Lemma	Input-Word	Lemma	Input-Word	Lemma
1.	abortion	abort	applicability	apply	annoying	annoy	allotments	allot
2.	abort	abort	appliers	applied apply	courteous	court	amazing	amaze
3.	abortive	abort	appliances	apply	creamery	cream	amazement	Amaze
4.	abruptness	abrupt	applications	apply	creamery	cream	amazed	amaze
5.	abscond	abscond abscond	applicants	apply	Curiosity	curios	American	america
6.	absent	absent	applicable	apply	daily	day	Amused	amuse
7.	absentia	absent	application	apply	death	death	Angrily	angry
8.	absenteeism	absentee absent	Appointee	appoint appoint	definitely	definite	compliance	comply
9.	absolute	absolute	approachable	approach	denial	denial	comprehension	comprehend
10.	absolutism	absolute	approaching	approach	departure	depart	conducting	conduct
11.	absorb	absorb	approached	approach	departure	depart	conduction	conduct
12.	absorbent	absorbent	argument	argue	deployable	deploy	conductive	conduct
13.	abstract	abstract	arrival	arrive	deployment	deploy	conductors	conduct
14.	absurdness	absurd	artistically	artistic artist	developments	develop	conductor	conduct
15.	absurd	absurd	astonish	astonish	developmental	development develop	conductivity	conduct
16.	absurdity	absurd	astromical	astromical	dictionary	dictate	conductivities	conductivity conduct
17.	absurdum	absurdum	astronaut	astronaut	dictation	dictate	conductress	conductress
18.	abundance	abund	astronomer	astronomy	dictator	dictate	compliance	comply
19.	abundant	abundant	astronomically	astronomy	dismissal	dismiss	comprehension	comprehend
20.	abuse	abuse	athletic	athlete	division	divide	conducting	conduct
21.	abuser	abuse	attended	attend	doable	do	conduction	conduct
22.	abusive	abuse	attentive	attend	donation	donate	conductive	conduct
23.	academicals	academic academy	attending	attend	easily	easy	conductors	conduct
24.	academically	academic academy	attention	attend	education	educate	conductor	conduct
25.	academy	academy academy	attractive	attract attract	educational	education educate	conductivity	conduct conduct
26.	academic	academy academy	attractive	attract	Employment	employ	conductivities	conductivity conduct
27.	academics	academy academy	attracted	attract	employee	employ	gracefully	graceful grace
28.	accent	accent	attracting	attract	employer	employ	happily	happy
29.	accentuate	accentuate	attraction	attract	employment	employ	hateful	hate
30.	accept	accept	australian	australia	enclosure	enclose	hopefully	hopeful hope
31.	acceptor	accept	awful	awe	engineer	engineer	humanitarian	human
32.	accessibly	access	bakery	baker	enviousness	envious envy	hydrate	hydrate
33.	access	access	basic	basic	envious	envy	idly	id
34.	accessory	access	beautiful	beauty	establishment	establish	identification	identify
35.	accessible	access	beggar	beggar	european	europe	immunise	immune
36.	accidental	accident	believable	believe	evasive	evasive	gracefully	graceful grace
37.	accomplishment	accomplish	birth	birth	excited	excite	successfully	success
38.	achievable	achieve	blackness	black	experience	experience	compliance	comply
39.	acoustically	acoustic	blacken	black	experiments	experiment	boyish	boy
40.	action	act	bravery	brave	experimental	experiment	preference	prefer

41.	active	active	breakage	break	explainable	explain	privacy	private
42.	actively	active	burglar	burgle			privatize	private
43.	activeness	active	busily	busy	expression	express	privatization	privatize private
44.	activity	activity	carbonate	carbonate	expressional	express	productive	product
45.	actor	actor	cheeriness	cheer	extraordinarily	extraordinary	profession	profess
46.	administrations	administrate	childish	child	fertilise	fertile	professor	profess
47.	admiration	admiration	chloride	chloride	flexible	flex	professional	profession profess
48.	admirable	admire	chosen	choose	foolish	fool	qualifications	qualify
49.	admired	admire	collector	collect	forcible	force	qualification	qualify
50.	admirer	admire	collection	collect	friendly	friend	rainy	rain
51.	admiring	admire	collision	collide	funny	fun	readable	ready
52.	advisable	advise	comical	comic	gently	gentle	realise	real
53.	algebra	algebra	communist	commune	government	govern	reconciliation	reconcile
54.	algebraic	algebra	complaints	complain	governable	govern	registrar	registry
55.	algerian	algeria	completely	complete	governed	govern	relaxed	relax
56.	asian	asia	complications	complex	governor	govern	preference	prefer
57.	algeria	algeria	compliance	comply	governmental	government govern	european	europe
58.	acoustic	acoustic	archaeological	archaeology	defective	defect	astronomical	astronomy
59.	acrobatic	acrobat	arguable	argue	detective	detect	attentive	attend
60.	admirable	admire	articulate	articulate	directive	direct	width	wide
61.	aesthetic	aesthete	aesthetical	aesthete	eleventh	eleven	length	length
Observation: **Number of words 282; Number of correct lemma 212** **Rate of Error: 2.3%**								

Table 4.17 shows the list of derived Adverb words with their Lemmas.

Table 4.17 Sample Data Set of Derived Adverb Words with Lemmas

	Input-Word	Lemma	Input-Word	Lemma	Input-Word	Lemma
1.	abnormally	abnormal	accurately	accurate	weakly	weak
2.	academically	academic academy	acoustically	acoustic	nobly	noble
3.	accidentally	accident	willingly	willing will	expertly	expert
4.	affectionately	affection	famously	famous	oddly	odd
5.	Angrily	angry	fashionably	fashion	officially	official office
6.	anxiously	anxious	freely	free	otherwise	otherwise
7.	Artfully	artful	furiously	furious fury	painfully	painful pain
8.	artistically	artistic artist	gently	gentle	personally	personal person
9.	awesomely	awesome	gracefully	graceful grace	politically	politics
10.	awkwardly	awkward	guiltily	guilty	possibly	possible
11.	badly	bad	happily	happy	probably	probable probe
12.	beautifully	beautiful beauty	harshly	harsh	proudly	proud
13.	briskly	brisk	helpfully	helpful help	punctually	punctual
14.	brutally	brutal	hopefully	hopeful hope	purposefully	purposeful purpose
15.	busily	busy	hurriedly	hurried hurry	quickly	quick
16.	calmly	calm	idly	idle	readily	ready
17.	capably	capable	imaginatively	imaginative imagine	really	real
18.	carefully	careful care	irresponsibly	irresponsible	regretfully	regretful regret
19.	cautiously	cautious	jealously	jealous	religiously	religious
20.	cheerfully	cheerful cheer	job-wise	job	rightly	right
21.	classically	classical class	jokingly	joking	romantically	romantic
22.	clearly	clear	joyfully	joyful	sadly	sad
23.	cleverly	clever	justly	just	safely	safe
24.	clockwise	clockwise	kindly	kind	secretly	secret
25.	colorfully	colorful color	knowledgeably	knowledge	silently	silent
26.	comfortably	comfort	lawfully	lawful	skillfully	skillful skill
27.	competitively	competitive	leisurely	leisure	sleepily	sleep

28.	completely	complete	lengthwise	lengthwise	steadily	steady
29.	confidently	confident	lifelessly	lifeless	suspiciously	suspicious
30.	counterclockwise	counterclockwise	lovingly	loving	tastefully	tasteful tasty
31.	cowardly	coward	loyally	loyal	tenderly	tender
32.	crazily	crazy	luckily	lucky	terribly	terrible
33.	customarily	customary custom	magically	magical magic	thoroughly	thorough
34.	definitely	definite	magnificently	magnificent	tragically	tragical tragic
35.	deliberately	deliberate	maturely	mature	uniquely	unique
36.	abnormally	abnormal	doubtfully	doubtful doubt	musically	musical music
37.	academically	academic academy	eagerly	eager	naturally	natural nature
38	universally	universe	vocally	vocal	warmly	warm
39	untruthfully	untruthful untruth	voluntarily	voluntary	watchfully	watchful watch
Observation: **Total number of Adverb words : 127 ; Number of lemma 127** **Rate of error: 1.5%**						

Comparative Analysis

The output generated by the proposed LemmaChase lemmatizer on the data set was compared with that of the output of the popular lemmatizers Like Stanford, Lemmagen, spaCy and WordNet lemmatizers. It was observed that LemmaChase identifies and generatesmore correct lemmas as compared to the lemmas generated by all others. It was seen that for the nominalized words, all other lemmatizers failed to find the lemma meanwhile LemmaChase would correctly generated the lemmas for most of the nominalized words. Table and Fig. Show this comparative analysis. Table 4.18 depicts the rate of error, precision, recall and F-score of LemmaChase and Stanford lemmatizers.

Table 4.18 Comparative Analysis between LemmaChase and Stanford Lemmatizer

LemmaChase Model	**Number of Words**	**Number of Lemma Generation**	**Number of Incorrect Lemma Generation**	**Rate of Error**	**Precision**	**Recall**	**F-Score**
Derived Words with Noun POS [11]	278	274	33	12%	245/(245+33)=0.88	1	0.94
Derived Words with Adjective POS[12]	175	173	26	15%	149/(149+26)=0.85	1	0.95
Derived Words with Adverb POS (Oxford Dictionary)	127	127	2	1.50%	125/(125+2)=0.98	1	0.98
1st Set of Derived and Inflected Words with Noun, ADJ, ADV POS (Penn University Resource , MorphoLex database)	282	212	6	2.30%	276/(276+6)=0.97	1	0.98
2nd Set of Derived and Inflected Words with Noun, ADJ, ADV POS (Penn University Resource, MorphoLex database)	418	123	24	5.90%	394/(394+24)=0.94	1	0.97
Stanford Lemmatizer	364	364	243	66.8%	121/(121+243)=0.33	1	2(0.33)/(0.33+1)=0.49

[11] Oxford Dictionary, https://usefulenglish.ru/writing/list-of-nouns-with-suffixes
[12] Oxford Dictionary, https://usefulenglish.ru/writing/list-of-derivative-adjectives

Precision = TP/ (TP+ FP) ; Recall = TP/(TP+ FN); F-Score=2(P*R)/(P+R)[13]

True positive (TP) indicates that the lemma which is identified by the model is correct.
False Positive (FP) indicates that the lemma which is identified by the model is not correct lemma.
False Negative (FN) indicates that the lemma which is actually correct is identified as an incorrect lemma.
True Negative (TN) indicates that incorrect lemma is identified as incorrect.
In this LemmaChase model, FN case is not applicable because model does not identify any extracted lemma as an incorrect lemma. Extracted lemma is always identified as correct by this model. So, FN is always 0 in case and calculated recall will always become 1 as per the formula of recall. For the LemmaChase model the following is the summary:

	Relevant Lemma	Non-Relevant Lemma
Retrieved Lemma	True Positives(TP) : Number of correct lemmas	False Positives (FP) : Number of incorrect lemmas
Not Retrieved lemma	False Negatives (FN) : NA	True Negatives (TN) : NA

Fig. 4.6 and Fig. 4.7 depict the graphical representation of Table 4.18.

[13] https://nlp.stanford.edu/IR-book/pdf/08eval.pdf

Fig 4.6 shows comparative bar diagram for LemmaChase and Stanford lemmatizers to depict the generation of incorrect lemma.

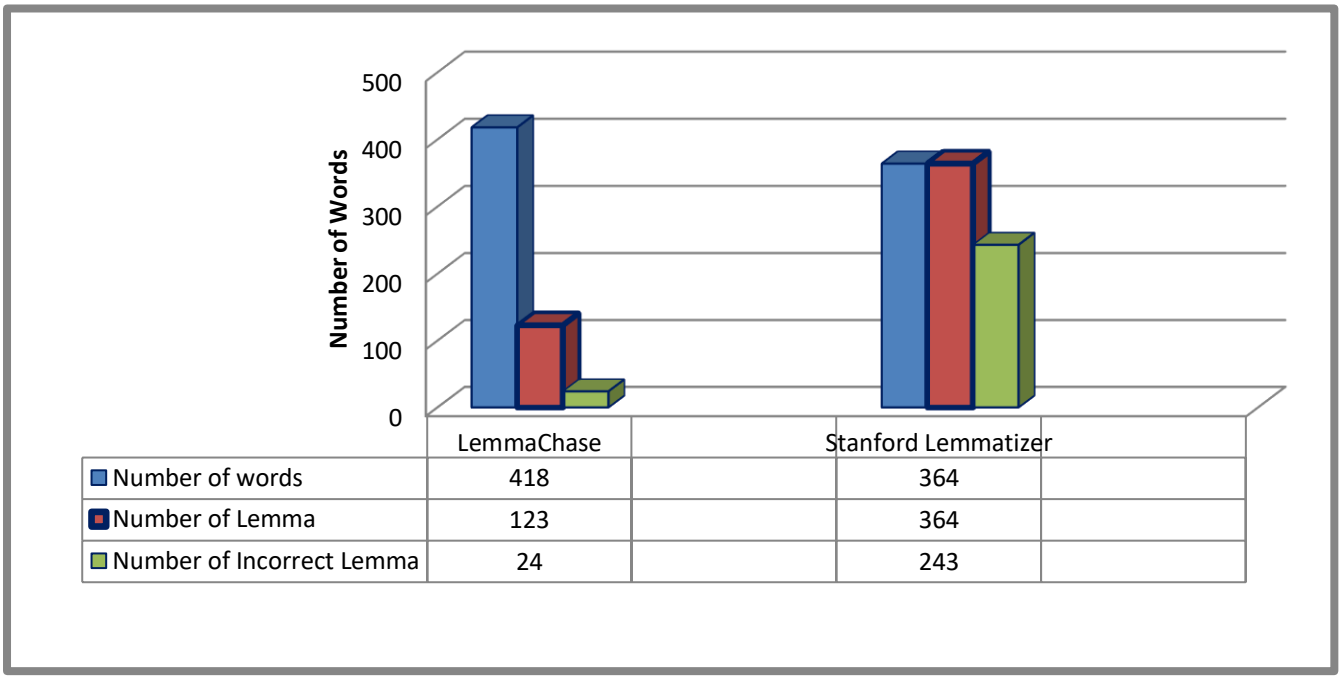


Figure 4.6 Bar Diagram for Incorrectness of LemmaChase and Stanford Lemmatizers

Fig. 4.7 shows Precision, F-Score of LemmaChase and Stanford Lemmatizer.

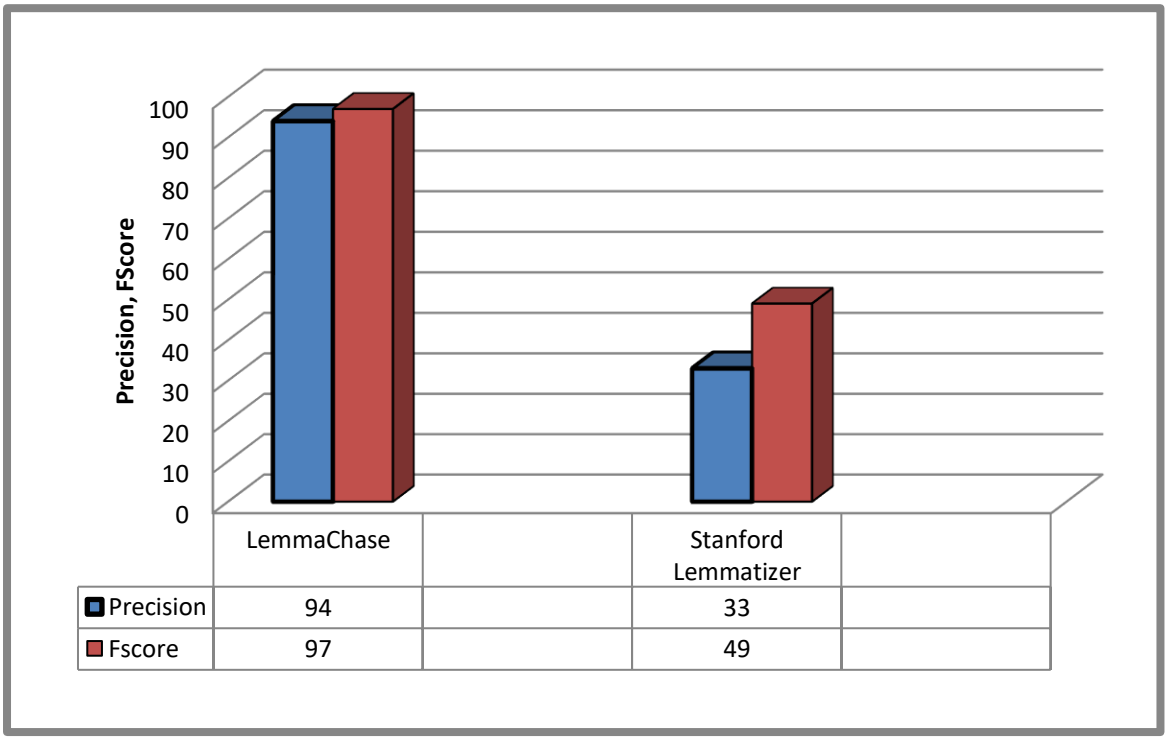


Figure 4.7 Bar Diagram for Precision and FScore of LemmaChase and Stanford Lemmatizers

Fig. 4.8 shows the numbers of words, number of lemma generation and number of incorrect lemma generation by LemmaChase.

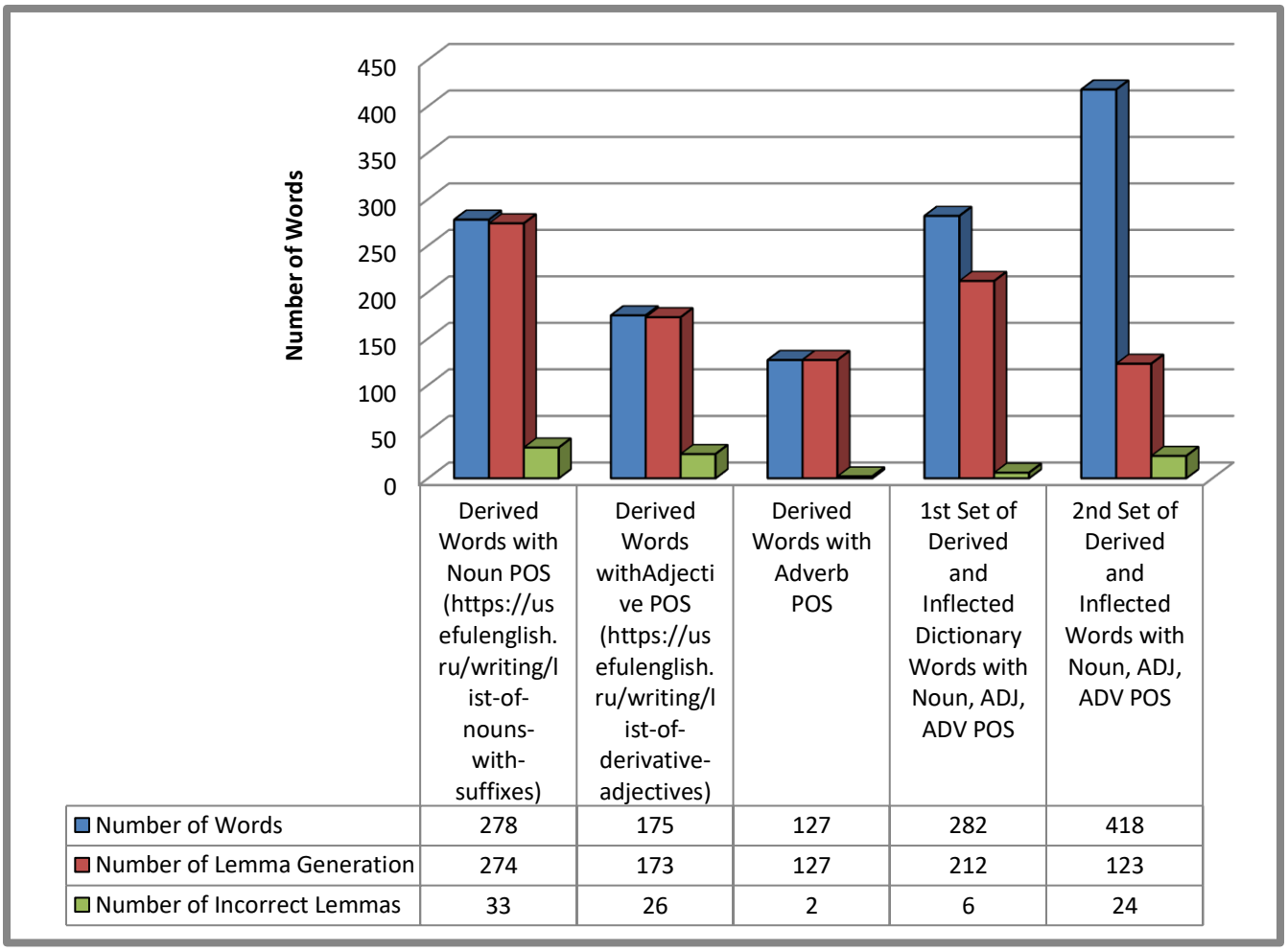


Figure 4.8 Bar Diagram of Incorrect Output by LemmaChase

Fig. 4.9 shows the Precision, Recall and F-score value of LemmaChase.

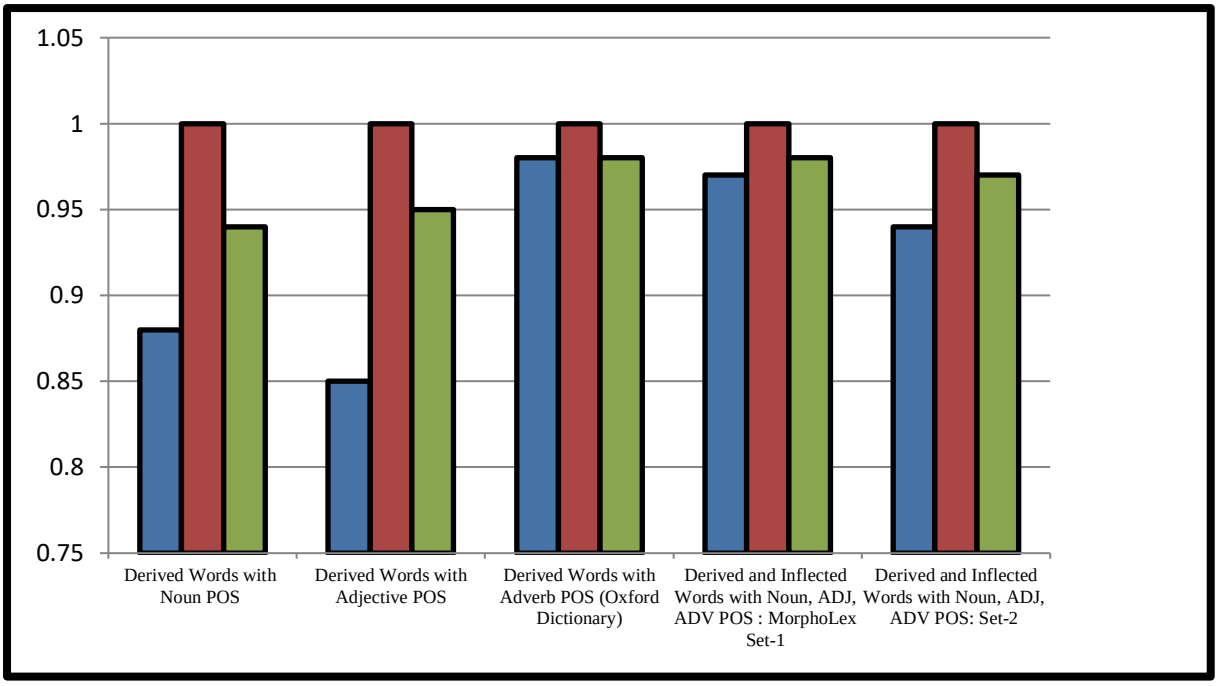
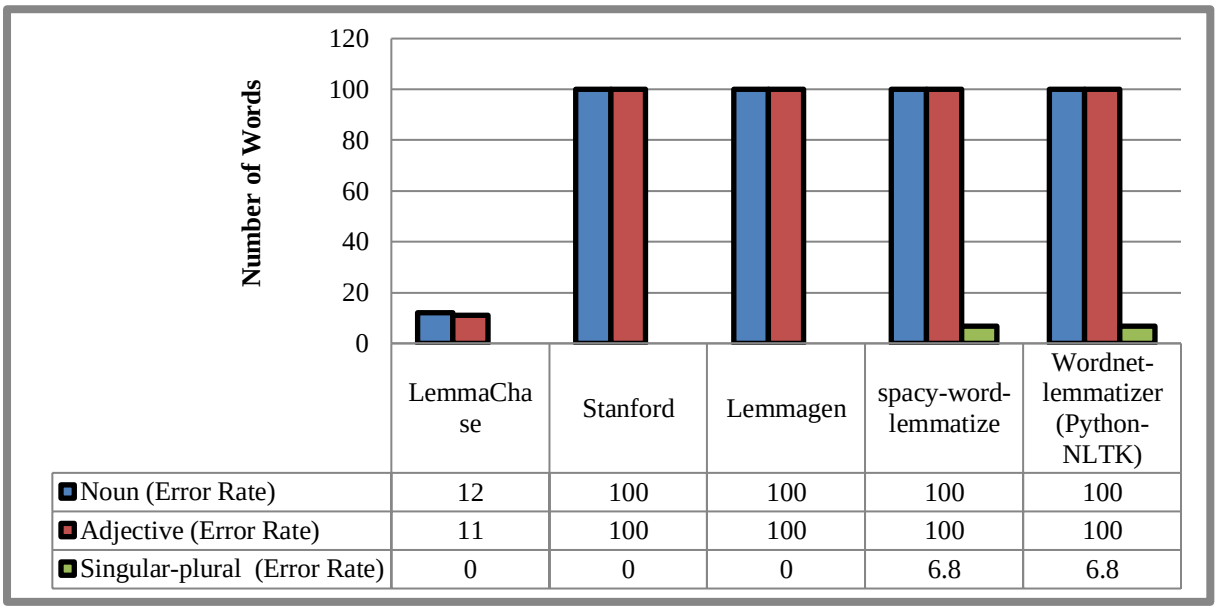


Figure 4.9 Bar Diagram for Precision, Recall and F-Score for LemmaChase

Fig. 4.10 shows comparative error-rate bar diagram of different Lematizers and LemmaChase in generation of lemmas from noun, adjective and irregular singular-plural words.



	LemmaChase	Stanford	Lemmagen	spacy-word-lemmatize	Wordnet-lemmatizer (Python-NLTK)
Noun (Error Rate)	12	100	100	100	100
Adjective (Error Rate)	11	100	100	100	100
Singular-plural (Error Rate)	0	0	0	6.8	6.8

Figure 4.10 Comparative Diagram of Error-Rate of LemmaChase with Existing Lemmatizers

Table 4.22 depicts incorrectness of LemmaChase and other lemmatizers.

Table: 4.22 Comparative Analysis in Lemma Generation

Lemmatizers	Rate of Errors in Percentage for Derived Words			
	Adjectives	**Nouns**	**Adverbs**	**Odd Nouns**
Stanford	100	100	100	100
Lemmagen	100	100	100	100
spaCy	100	100	100	100
WordNet- (Python-NLTK)	100	100	100	100
LemmaChase	**11**	**12**	**1.5**	**15**

Note: Odd Nouns are nouns originated from different lemmas e.g. universe, university, statement, etc.

4.8 Conclusion and Summary

It is observed that the proposed lemmatizer-LemmaChase handles maximum number of nominalized words which are not handled by available popular lemmatizers. It handles most of the morphed and derived words (singular, plural in Noun, irregular singular-plural/ verbs /adjective/ adverbs) with any POS form in a text. In this proposed model, each morphed word whether related to the same lemma or different lemma are processed individually which leads to consuming more processor time. When words “application”, “applications”, “applicant”, “applicants”, “applicable” and “applicability” are available in a text, all these six allied words are parsed and processed individually to generate the lemma “apply”. The double or triple suffixation words (deployable, applicability, imaginatively, regretfully) are still not accurately processed by LemmaChase model to generate their corresponding lemmas. These words are not available in WordNet dictionary also.

This model has been published as a paper titled “LemmaChase: A Lemmatizer” in International Journal on Emerging Technologies 11(2): 817-824(2020).

To overcome the limitations of LemmaChase, another model LemmaQuest is proposed which is discussed in the next chapter.