Research review of Rule Based Gujarati Grammar Implementation with the Concepts of Natural Language Processing (NLP)

Mr. Nitesh G. Patel, Ph.D. Scholar
Department Of Computer Science, Gujarat Vidyapith, Ahmadabad, Gujarat
nitesh.mscit@gmail.com

Dr. Dhiren B. Patel, Professor
Department Of Computer Science, Gujarat Vidyapith, Ahmadabad, Gujarat
dhiren_b_patel@gujaratvidyapith.org

ABSTRACT

Natural language processing which is abbreviate as NLP. It is a prominent area of research now days. NLP's research and application discovers how computer can be used to understand and manipulate natural language's speech & text to do some interesting and useful things. The term language in Natural Language Processing (NLP) has to be understood as natural languages like (Gujarati, Hindi, English etc...).
The center of attention of this paper is to get the review of available technology and its proper method to apply the some of the aspects of NLP for the implementation of Gujarati Grammar. Our main focus is on the Rule base implementation because similar to every Indian scripting language Gujarati also have its own specified rules of composition for combining the consonants, vowels and modifiers.

KEYWORDS

Natural Language Processing, Gujarati Language, grammar, Stemming, Rule Based, Morphology

INTRODUCTION

In India there are various peoples talking & communicating in various languages and huge literature is available in different local languages which are not understandable to others in India itself.

Here, we are concentrating on Gujarati Language. Gujarati language is belongs to Indo-Aryan language and Indo-European language family and it is also closely related to 'Hindi' which is India's most used language. Gujarati is the official language of the Gujarat which is India's westernmost state. Approximately 50 million people speak Gujarati language in India and near about 1 million people speak outside of India also as it is 23rd most widely used language in the whole world. The Gujarati phoneme set mainly consists of twenty-four consonants and eight vowels.

અ	આ	ઇ	ઈ	ઉ	ઊ	ઋ	એ	ઐ	ઓ	ઔ	Initial
a	ā	i	ī	u	ū	ṛ	e	ai	o	au	Vowels
[ə]	[a]	[i]	[i]	[u]	[u]	[ɾi]	[e/ɛ]	[əj]	[o/ɔ]	[əw]	
ક	ખ	ગ	ઘ	ઙ							Velar
ka	kha	ga	gha	ṅa							
[kə]	[kʰə]	[gə]	[gʰə]	[nə]							
ચ	છ	જ	ઝ	ઞ							Palatal
ca	cha	ja	jha	ña							
[tʃə]	[tʃʰə]	[dʒə]	[dʒʰə]	[nə]							
ટ	ઠ	ડ	ઢ	ણ							Retroflex
ṭa	ṭha	ḍa	ḍha	ṇa							
[tə]	[tʰə]	[də]	[dʰə]	[nə]							
ત	થ	દ	ધ	ન							Dental
ta	tha	da	dha	na							
[tə]	[tʰə]	[də]	[dʰə]	[nə]							
પ	ફ	બ	ભ	મ							Labial
pa	pha	ba	bha	ma							
[pə]	[pʰə]	[bə]	[bʰə]	[mə]							
ય	ર	લ	વ								Glide and Liquid
ya	ra	la	va								
[jə]	[rə]	[lə]	[wə]								
શ	ષ	સ	હ	ળ	ક્ષ	જ્ઞ					Fricative & Other
śa	ṣa	sa	ha	ḷa	kṣa	jña					
[ʃə]	[ʃə]	[sə]	[ɦə]	[ɭə]	[kʃə]	[dʒnə]					

Fig 1: Gujarati phoneme set (source: Internet)

Natural language processing is the branch of Computer Science with two basic goals:

1. Understanding grammar and rules to work upon the specified Natural language.
2. Build the system that analyzes the Natural language from different aspects and minimized the man-machine gap.

On the basis of theory there are two basic aspects of NLP:

1. Natural Language Understanding (NLU)
    - Lexical ambiguity
    - Syntactical ambiguity
    - Referential ambiguity
2. Natural Language Generation (NLG)
    - Text Planning
    - Sentence Planning
    - Text Realization or understanding

A review of the various methods and paradigms of NLP with respect to the already mentioned four criteria related to the specification of:

- Syntax and semantics,
- Learn-ability,
- Computational complexity and
- Ambiguity resolution

METHODOLOGY:

From the below list out methodologies we have to take decision that which is more important and useful for implementing the grammatical rules of Gujarati using NLP. This task will lead us to the expected solution of our goal.

- Lexical Analysis: It Deals with recognition and identification of structure of the sentence. It divides the paragraph into sentences, phrases & words. It doesn't deal with meaning of words, sentences & phrases.

- Syntactic Analysis: It is mainly relies on grammar of sentence which analyzed in order to get the relationship among different words in sentence. Here the sentence is parsed as Adjectives, Noun, Verbs & other part of sentences.

- Semantic Analysis: The actual meaning of the sentence is extracted from words used in it. It checks weather the word individually or group generate any meaning or not.

- Disclosure Integration: In disclosure Integration the meaning of sentence is verified with sentence which comes before it. So, sentences can relate with each other for proper meaning rather than individual meaning.

- Pragmatic Analysis: Here the sentences are re-interpreted to verify the correctness of meaning in particular given context or situation. The Real world knowledge of language is must required.

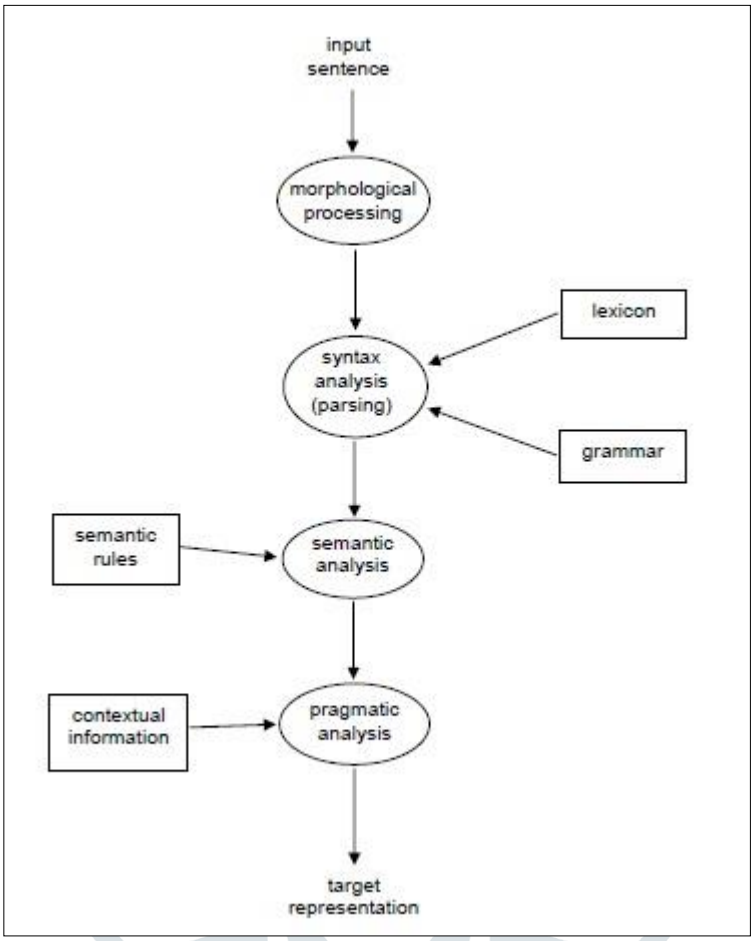

Fig 2: Process of NLP (source: Internet)

REVIEW OF LITRATURE

We have reviewed multiple research papers with a good amount of variation of technology & methodology of NLP. We have reviewed the work done in mainly Gujarati language because our major focus is on the implementation of the path-way of Gujarati Grammar in this digital world.

The research paper titled "Morphological Rule Set and Lexicon of Gujarati Grammar: A Linguistics Approach" presents morphological rules for Gujarati language classes and lexicon database. In this paper we have presented useful package composed of morphological grammar rules, dictionary, test data, and a set of API. The rules are implemented in database for further processing and development of morphological analyzer for Gujarati language.[1]

The research paper titled "A Lightweight Stemmer for Gujarati shown an implementation of a rule based stemmer of Gujarati". They have shown the creation of rules for stemming and the richness in morphology that Gujarati possesses. The stemmer is able to capture most of the morphological variants. They have like to perform a more rigorous error analysis so that a detailed error analysis report can be provided.[2]

The approach proposed by the authors in "Gujarati Phonetics and Levenshtein based String Similarity Measure for Gujarati Language" paper is capable enough to identify similar strings of Gujarati language which are phonetically similar and are different because of just similar sounding matras or vowels. The work can be further tested on relevant corpus. The approach can be improved by improving the stemmer.[3]

The authors have developed "Saaraansh: Gujarati Text Summarization System". Though human generated summaries are difficult to achieve by automatic text summarization, still using linguistic components like Stemmer and String similarity measure, summary with good recall can be achieved. The performance of text summarization improves by adding linguistics components to it.[4]

The same authors have also developed "Dhiya: A stemmer for morphological level analysis of Gujarati language". In this paper, authors suggest DHIYA a stemmer for Gujarati language. This stemmer is based on the morphology of Gujarati language. To develop the stemmer, inflections which appeared most in Gujarati text were identified.[5]

To understand a language, analysis has to be done at word level, sentence level, context level and discourse level. Morphological analysis comes at the base of all, as it is the first step to understand a given sentence.[6]

In The paper "Hybrid Inflectional Stemmer and Rule-based Derivational Stemmer for Gujarati" they have used two stemmers for Gujarati- a lightweight inflectional stemmer based on a hybrid approach and a heavyweight derivational stemmer based on a rule-based approach.[7]

IMPORTANT ASPECTS OF NLP

- Normalization and tagging: acronyms can be specified as "□□.□." or "□□□" so these should be tagged and normalized.

- Lemmatization: Lemmatization uses a language dictionary to perform an accurate reduction to root words. Lemmatization is strongly preferred to stemming if available.[22]

- POS (Part-Of-Speech) tagging: POS is the task of assigning each word its (POS) tag. The most common ones are noun, verb, determiner, adjective and adverb.[22]
  Methods for tagging multiple methods available:
  Default tagger, Regular expression tagger, unigram tagger and n-gram taggers.

- Named-entity recognition (NER) : NER is a sub part of information extraction that use to find and classify the named entities in textual data into pre-defined categories like the names of people, companies, locations, expressions, quantities etc.[22]

- Stemming: Stemming uses simple pattern matching and after that it will simply strip suffixes of tokens or word. In short it will Search the root of the particular word. [22] (Example: □□□□□□ ------ > □□□□)

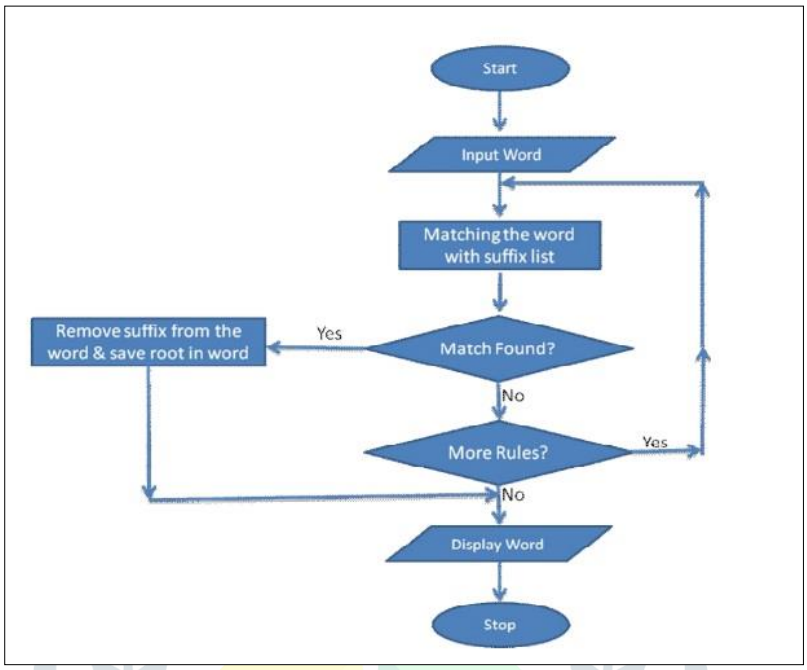

Fig 3: Process of Stemming (source: Internet)

RESEARCH CHALLENGES FOR FUTURE WORK

- Search for the standardization in the well known Indic language Gujarati & search for the proper solution for implementation.

- Implement the techniques and methodology which are new and not implemented yet & search for the loopholes and limitation of it.

- Willing to design a proper path or roadmap for complete Gujarati grammar implementation with suggestive design.

- To resolve some of the basic problems of implementation like :
  - Ambiguity in translation & semantics of each character, word & line.
  - Lack of knowledge or vocabulary of Gujarati language.
  - Hard to check the proper Transition to another language for cross checking.

CONCLUSION

In this effort of paper our main focus is on the "how Gujarati grammar's implementation can be possible using available technology and suitable concepts of NLP?",like other languages' implementation. In Reality, the implementation of the grammar of Gujarati language seeks a huge work effort. The Biggest problem of implementation is 'standardization' of the Gujarati Language and its Grammar. Here we try to show at least a proper roadmap for the some of the language based rules in process of NLP. In future any researcher may apply rule based morphological approach for Gujarati language classification and extend this work by implementation of grammar rules which can be used in other research in area of Natural Language Processing.

REFERENCES

[1] Kapadia Utkarsh N. , Desai Apurva A., Morphological Rule Set and Lexicon of Gujarati Grammar: A Linguistics Approach, Vnsgu Journal Of Science And Technology, Vol.4. No. 1, July, 2015 127 - 133,ISSN : 0975-5446

[2] Juhi Ameta, Nisheeth Joshi, Iti Mathur, A Lightweight Stemmer for Gujarati, Banasthali University, Rajasthan, India

[3] Jikitsha R. Sheth, Bankim C. Patel, Gujarati Phonetics and Levenshtein based String Similarity Measure for Gujarati Language, Conference Paper - February 2015, https://www.researchgate.net/publication/314153559

[4] Jikitsha R. Sheth, Bankim C. Patel, Stemming Techniques and Naïve Approach for Gujarati Stemmer ,International Conference in Recent Trends in Information Technology and Computer Science (ICRTITCS - 2012) Proceedings published in International Journal of Computer Applications® (IJCA) (0975 – 8887)

[5] Jikitsha R. Sheth, Bankim C. Patel, Saaraansh: Gujarati Text Summarization System ,IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol.7, No.3, May-June 2017

[6] Jikitsha R. Sheth, Bankim C. Patel, Dhiya: A stemmer for morphological level analysis of Gujarati language, Conference Paper · February 2014, DOI: 10.1109/ ICICICT.2014.6781269

[7] Kartik Suba(DDU), Dipti Jiandani(DDU),Pushpak Bhattacharyya(IIT Bombay), Hybrid Inflectional Stemmer and Rule-based Derivational Stemmer for Gujarati, South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011

[8] Harshali B. Patil, B. V. Pawar, and Ajay S. Patil , A Comprehensive Analysis Of Stemmers Available For Indic Languages, International Journal on Natural Language Computing (IJNLC) Vol. 5, No.1, February 2016 DOI: 10.5121/ijnlc.2016.5104 45

[9] Anil Kumar Singh , A Computational Phonetic Model for Indian Language Scripts ,Language Technologies Research Centre IIIT, Hyderabad, India)

[10] Alok Kumar, Saurabh Sharma, Mushahid Raza, Lexical Analysis of Devanagari Hindi Language, International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 6 Issue 7 July 2017, Page No. 21916-21921 Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i7.05)

[11] R. Vijaya Lakshmi, Dr. S. Britto Ramesh Kumar , Literature Review: Stemming Algorithms for Indian and Non-Indian Languages ,International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014) Vol. 2, Issue 3 (July - Sept. 2014) ISSN : 2347 - 8446 (Online)

[12] Prof. Langote Manojkumar,Miss Kulkarni Sweta, Miss Mansuri Shabnam, Miss Pawar Ankita and Miss Bhoknal Kishor, Role of NLP in Indian Regional Languages, IBMRD's Journal of Management and Research, Print ISSN: 2277-7830, Online ISSN: 2348-5922, September 2014)

[13] Matic Perovšek, Janez Kranjc, Tomaž Erjavec, Bojan Cestnik, Nada Lavra, TextFlows: A visual programming platform for text mining and natural language processing ,5 January 2016

[14] Yoav Goldberg , A Primer on Neural Network Models for Natural Language Processing, Journal of Artificial Intelligence Research 57 (2016) 345{420 Submitted 9/15; published November 2016

[15] Djalma Padovani, João José Neto , Adaptive Automata Applied to Natural Language Processing, ScienceDirect Procedia Computer Science 109C (2017) 1152–1157, International Workshop on Adaptive Technology (WAT 2017)

[16] Ms. Rijukta Pathak, Mr. Biju Thankachan, Natural language processing approaches, application and limitations ,International Journal of Engineering Research & Technology(IJERT) Vol. 1 Issue 7, September – 2012 ISSN: 2278-0181

[17] Ekansh Gupta ,Rohit Gupta, Mentored by: Prof. Amitabha Mukerjee , Natural Language Processing Hindi ↔ English Parallel Corpus Generation from Comparable Corpora for Neural Machine Translation ,IIT KANPUR

[18] Hasan Kabir, Syed Raza Shahid, Abdul Mannan Saleem and Sarmad Hussain, Natural Language Processing for Urdu TTS System

[19] Dima Suleimana,Arafat Awajana,Wael Al Etaiwia, The Use of Hidden Markov model in natural Arabic language Processing: a Survey, International conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017)

[20] A. F. Alajmi, E. M. Saad and M. H. Awadalla, Hidden markov model based Arabic morphological analyzer ,Communication and Electronics Department, Faculty of Engineering, Helwan University, Egypt. Accepted 24 February, 2011

[21] Juhi Ameta, Nisheeth Joshi and Iti Mathur, Improving The Quality Of Gujarati-Hindi Machine Translation Through Part-Of-Speech Tagging And Stemmer Assisted Transliteration, International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3, June 2013

[22] Other resources : Web link : Wikipedia- https://en.wikipedia.org/wiki/Natural-language_processing

[23] Other resources: PDF book : “Natural language processing” Book by ELA Kumar