Semantic Analysis of Answering Questions from the Tamil Text Document

S.ArulJothi[1],

[1]AssistantProfessor, Department of Computer Science and Engineering Nadar Saraswathi College of Engineering and Technology,Theni,Tamilnadu

M.Ambika[2] **A.Gowmari**[3] **M.Monisha**[4]
UG Students

[2,3,4] Department of Computer Science and Engineering Nadar Saraswathi College of Engineering and Technology,Theni,Tamilnadu

Abstract— **"**Natural language processing system" is a branch of AI that helps computer to understand, interpret and manipulate and respond to human in their natural language. NLP is a subfield of computer science that deals with AI. Linguistic phrases and sentence can be formed with words, intuitions about well-formed and meaning, the curbs the possible meaning for a sentence in a mathematical model of structure. The question answering is one of the applications involved in NLP. The main aim is to develop AI system, to understand the valid Tamil sentence which follows the grammatical rules and constraints imposed by the language and answer the question by semantic analyses of POS tagging.

Keywords-NLP, Part-of-Speech Tag

I.INTRODUCTION

Within the past decade, address replying (QA) capability in look motors and past has advanced as an imperative device to coordinate the more requesting and particular data needs of clients who are communicating their needs as questions rather than fair catchphrase questions, and they are not substance with a returned list of records to filter through. We have created AI in NLP the most reason of us extend is programmed address generator for Tamil, could be a dialect handling framework which is used to produce address for substantial Tamil sentence which takes after the syntactic rules and limitations forced by the language. Natural dialect handling (NLP) may be a subfield of etymological, computer science, and manufactured insights concerned with the intuitive between computers and human dialect NLP is the computerized approach to examining content that's based on both a set of hypotheses and a set of innovations. Since content can contain data at numerous distinctive level, from basic word or token-based representations, to wealthy progressive syntactic representations, to high-level consistent representations over record collections. 1) First completely reusable content collection for Tamil content report QA on the content book verses that straightforwardly reply the address were comprehensively extricated and annotated.2) To encourage the utilize of in assessing content collection archive we propose a few assessment measures to bolster the diverse sorts of questions and nature of verse-based answers whereas coordination the concept of fractional coordinating of answers within the evaluation.

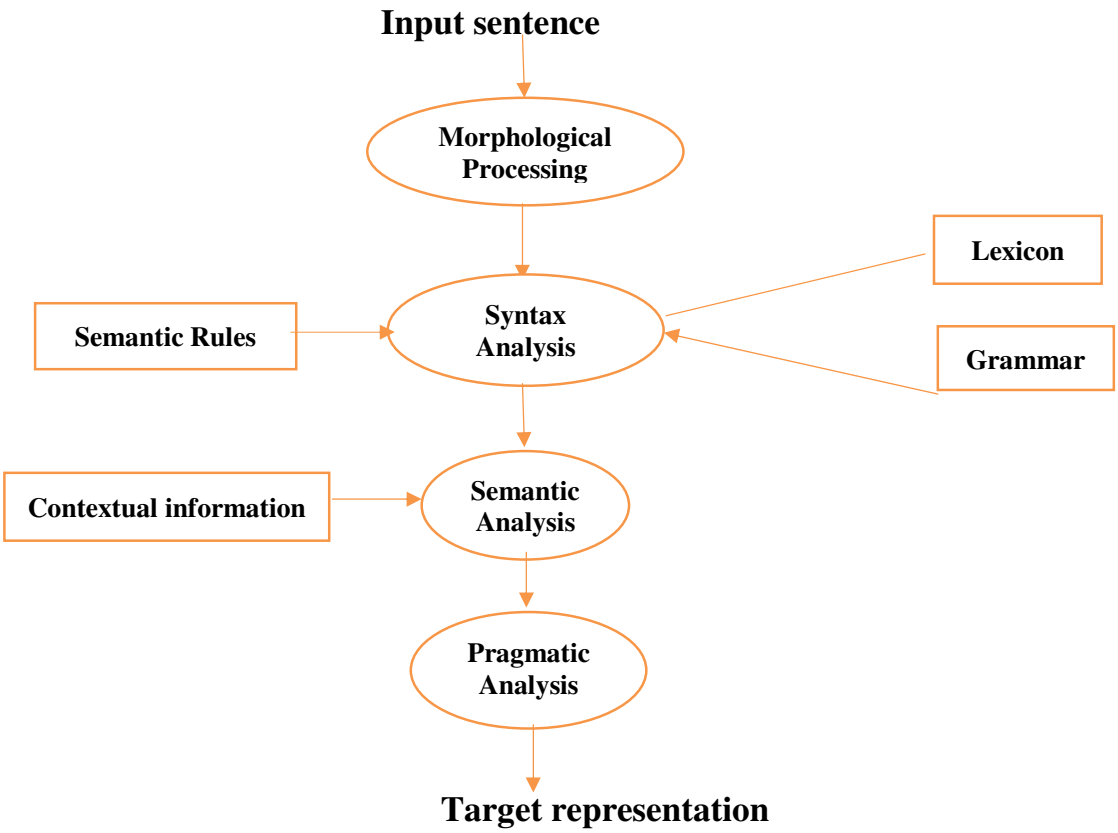

Fig.1 Phases or logical steps

II.EXISTING SYSTEM

Programmed era of normal dialect content is a basic errand in numerous dialect preparing like summarization, programmed record era, address replying framework, interpreter etc. It comprises of a learner to memorize how to realize a sentence from the substance of semantic part data. This learner is to be outlined as a factual demonstrate that's defined from a preprocessed corpus of sentences. This preprocessing is dealt with by pos labeling, chunking and semantic part labeling. A POS labeling instrument [1] for Tamil is created in Anna University. Similarly Chunking apparatus [2] for Tamil is additionally created at Anna College which can too been utilized for Chunking of POS labeled writings. These frameworks serve to be the pre runner/ existing works that comply with us extend.

III. PROPOSED SYSTEM

INPUT:

The input to this dialect preparing framework could be a substantial Tamil sentence which complies all the auxiliary and linguistic limitations of that dialect. For case the input may be a sentence of this kind.

TAMIL POS TAGGER:

This portion of the language preparing framework labels each of the words on the given input sentence with appropriate descriptor labels which empowers the era of questions naturally with the assistance of those labels. The method of labelling and the labels utilized are portrayed underneath.

NOUN MARKER TAG:

Within the handle of labelling the thing marker each and each word of the sentences is labelled as the thing at first and after completion of the labelling method by coordinating the semantic and basic rules of the language the remaining untagged words are considered words are considered to be thing. Therefore, above sentence are explained:"கீதா தினந்தோறும் பள்ளிக்கு செல்லும் வழியில் ஒரு வயதான பெரியவரைக் காண்பாள் , பின்பு ஒரு நாள் அவருக்கு உணவு கொடுத்து உதவினாள்"

VERB MARKER TAG:

In this of labelling the verb marker the taking after sub labels are too made hence empowering the era of address which is however to happen within the afterward stage. The sub labelling done which helps the forms of verb recognizable proof is described underneath.

GENDER MARKER TAG:

Each word of the input Tamil sentence is checked for the nearness of Gender markers such as

இவன்,உவன்,அவன்,எவன்,அவள் / இவள்,அவர்கள் / இவர்கள், at the conclusion of the word which describes male, female, plural and sexual orientations other than male or female. In the event that such gender marker labels are found within the word at that point the calculation continues in finding the tense markers which is depicted underneath.

TENSE MARKER TAG:

A tense marker shows up fair underneath the sexual orientation marker and is utilized to conclude that the word with a tense marker gone before by a sexual orientation marker is certainly a verb.

Present tense: **கிறு,கின்று,ஆநின்று**

Past tense: **த், ட், ற்,இன்**

Future tense: **ப்,வ்**

Thus a word is tagged as a verb if it is of the form:

Verb=action word + tense marker + gender marker

தினந்தோறும்= தினம்+தோறும்

TIME MARKER TAG:

The word is checked against the taking after predefined time words such as **இன்றுநேற்று, நாளை**markers such as date and week markers and in case these marker words are display the word gets labeled with a **நாளை** time marker descriptor tag.

NOUN DESCRIPTOR TAG:

For tagging a word as a noun descriptor the word has to be checked against these following rules. The word must end with the rhyme and word which succeeds this word must be a noun.

VERB DESCRIPTOR TAG:

For labelling a word as a thing descriptor the word has got to be checked against these taking after rules. The word must conclusion with the rhyme and word which succeeds this word must be a thing.

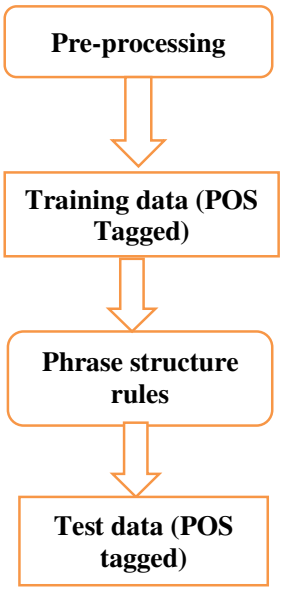

Fig1.1 Overview of proposed system model 1

QUESTION ANALYSER:

The stage of the dialect preparing framework gets the labelled sentence as the input and produces all conceivable questions as the output. The fitting labels are utilized to create questions by taking after the rules given underneath the labelled sentence is once more filtered from cleared out to right word by word and each time a word is supplanted within the sentence based on fitting labels to that word to create address:

RULES FOR QUESTION ANALYSER:

Rule 1: Replace the time marker tag in the sentence with the word to **"எங்கு"**generate a question. Considering the example, the question generated by replacing the time marker will be of the form

கீதா தினந்தோறும் எங்கு செல்வாள் ?

Rule 2: Replace the noun quantifier tag in the sentence with the word "**என்ன"** to generate a question.

Rule 3: Replace the first occurrence of noun tag in the sentence with the word "யாருக்கு" to generate a question if the gender is either male or female else replace the word "யாரைக்" with considering the example the question generated by replacing the time marker will be of the form

கீதா யாருக்கு உணவு கொடுத்து உதவினாள்?

Rule4: Replace the verb quantifier tag in the sentence with the word "**என்ன**" and also delete the noun and its corresponding noun quantifier to generate a question. Considering the example, the question generated by replacing the time marker will be of the form

Rule 5: Find the words tagged as nouns and ending with the rhymes ஐ ,ஆள், க்கு, இன், அது, கண் and replace them with and replace them with '**யாரை யாரால் யாடுகக்கு யாரின் யாரது யார்கண்**' if the verb in that sentence is either has either male or female gender descriptor else replace them with '**எதை எதால் எதற்கு எதின்**'.

Note: The second question generated by following the rule 5 is not valid which can be detected by checking the structural pattern of the language. (Check 5. Future work portions for further reference regarding this issue)

Thus the various questions generated by following the above rules for the given input sentence are as follows:

1, கீதா தினந்தோறும் எங்கு செல்வாள்?

2, கீதா பள்ளிக்கு செல்லும் வழியில் யாரைக் கண்டாள் ?

3, கீதா யாருக்கு உணவு கொடுத்து உதவினாள்?

4, கீதா அந்த பெரியவருக்கு என்ன உதவி செய்தாள்?

5, கீதா தினந்தோறும் எங்கு செல்லும் வழியில் அந்த பெரியவரைக் கண்டாள்?

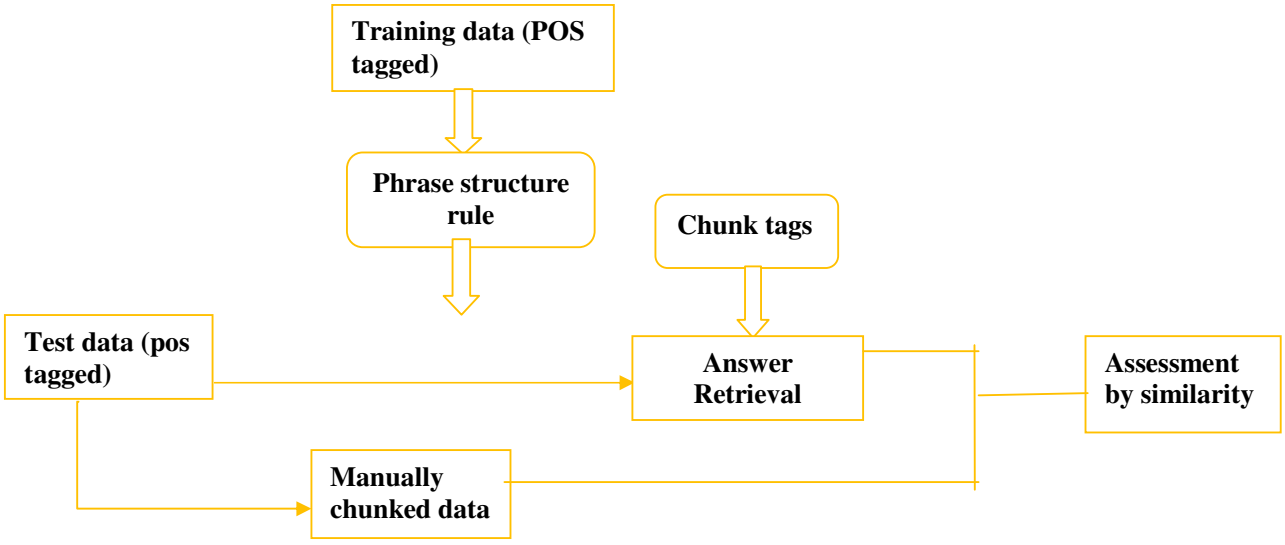

Fig 1.2Overview of proposed system model 2

ALGORITHM

```
Algo QuestionGen ()
Begin
        Noun Count=0;
        While (is Next Word ()! =NULL) do
         Begin
//Tamil Pos Tagging ()
          Tag each word as noun;
           If (is Gender Descriptor () and id Tense Descriptor ()) then
                    Tag word as verb;
           Else if (is Rhyme End("அ") and is next word () =noun) then
  Tag word as noun quantifier
Else if (is Rhyme End ("அ") or is Rhyme end    ("இ") then
             Tag word as verb quantifier;
         Else if (is time descriptors available ()) then
             Tag word as time descriptor;
End

While (is next word ()! = Null) do
Begin
//question generation ()
            If (is noun tag () and noun count == 0) then
                        If (is male () or is female ()) then
             Replace word with "யாருக்கு";
        Else
                        Replace word with "□□□□□□";
           Noun count++;
If (is time tag ()) then
         Replace word with "எங்கு";
Noun count++;
If (is time tag ()) then
         Replace word with "என்ன";
If (is noun quantifier ()) then
                        Replace word with "என்ன";
                         Remove (noun Quantifier (previous word ()));
                         Remove (previous word ());
                        End

                      End
```

FUTURE WORK

This address era framework which we have planned permits supplanting as it were a word within the input sentence with one address word but it takes after the auxiliary and linguistic rightness of the dialect. This effective address era may be made strides by permitting more than one substitutions of labelled words in a sentence with address words and creating ideal questions by keeping an upper and lower boundary on number of substitutions allowed. Moreover, classification of things based on living and non-living things makes a difference us to superior get it the thing quantifier and extra data can be given to the thing descriptor and what it right now evaluates can be recognized and era of comparing questions can be made conceivable.

CONCLUSION

In this way this programmed address era framework executed over can be made utilize for surrounding questions given a substantial input sentence or gather of sentences. This dialect preparing framework finds application in producing ideal questions which helps in way better understanding of auxiliary and syntactic angles of a dialect. This same strategy can be connected to other dialect as it were changing the essential pieces of grammar for that comparing language.

REFERENCES:

[1] Chowdhury, Gobinda G. "Natural language processing." Annual review of information science and technology 37.1 (2003): 51-89.
[2] Liddy, Elizabeth D. "Natural language processing." (2001).
[3] Spyns, P. "Natural language processing." Methods of information in medicine35.4 (1996): 285-301.
[4] Grishman, Ralph. "Natural language processing." Journal of the Association for Information Science and Technology 35.5 (1984): 291-296.
[5] Hirschman, Lynette, and Robert Gaizauskas. "Natural language question answering: the view from here." natural language engineering 7.4 (2001): 275-300.
[6] Burke, Robin D., et al. "Question answering from frequently asked question files: Experiences with the faq finder system." AI maga- zine 18.2 (1997): 57.
[7] Lehnert, Wendy G. Strategies for natural language processing. Psychology Press, 2014.
[8] Covington, Michael A. Natural language processing for Prolog pro- grammars. Englewood Cliffs (NJ): Prentice hall, 1994.
[9] Lehnert, Wendy G. The Process of Question Answering. No. RR-88. YALE UNIV NEW HAVEN CONN DEPT OF COMPUTER SCIENCE, 1977.
[10] Moldovan, Dan, et al. "The structure and performance of an open- domain question answering system." Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2000.
[11] Sowa, John F. Knowledge representation: logical, philosophical, and computational foundations. Vol. 13. Pacific Grove: Brooks/Cole, 2000.
[12] M. Alrabiah, A. Al-Salman, E. S. Atwell, and Nawal Alhelewh. 2014. KSUCCA: A key to exploring Arabic historical linguistics. International Journal of Computational Linguistics 5, 2 (2014),27–36.
[13] Eric Atwell, Nizar Habash, Bill Louw, Bayan Abu Shawar, Tony McEnery, Wajdi Zaghouani, and Mahmoud El-Haj. 2010. Understanding the Quran: Anewgrand challenge for computer science and artificial intelligence.In Proceedings of the Conference on Grand Challenges in Computing Research(GCCR'10).
[14]YonatanBelinkov,AlexanderMagidow,AlbertoBarrón-Cedeño, AviShmidman, and MaximRomanov.2019. Studying the history of the Arabic language : Language technology and a large-scale historical corpus. Language Resources and Evaluation 53 (2019),771–805.
[15] Hoa Trang Dang, Diane Kelly, and Jimmy Lin. 2007. Overview of the TREC 2007 question answering track. In Proceedings of the 15th Text Retrieval Conference(TREC'07).

[16] Hoa Trang Dang, Jimmy Lin, and Diane Kelly. 2006. Overview of the TREC 2006 question answering track. In Proceedings of the 14th Text Retrieval Conference(TREC'06).
[17] Aimad Hakkoum and SaidRaghay.2016.Semantic Q&A system on the Quran. Arabian Journal for Science

and Engineering 41, 12 (Dec. 2016), 5205–5214.
[18]M.A.HamdelsayedandE.S.Atwell.2016.Islamic applications of automatic question-answering. Journal of Engineering and Computer Science 17, 2 (2016), 51–57.
[19]Mohamed Adany Hamde lsayedand E.S.Atwell.2016. Using Arabic numbers (singular,dual,andplurals) patterns to enhance question answering system results. In Proceedings of the 4th International Conference on Islamic Applications in Computer Science and Technologies(IMAN'16).
[20] Mohamed Adany Hamdelsayed, Ebtihal Mustafa Elamin Mohamed, MohamedAlmoayed TajAlsir Mohamed Saeed, Akbar Musa Ai, Edress Babiker Edress Mohamed Mhmoud, Maha Ali Mahmoud, Ahmed Shamat, and Eric Atwell. 2017. Islamic application of question answering systems: Comparative study. Journal of Advanced Computer Science and Technology Research 7, 1 (2017), 29–41.