MORPHEME MASTERY FOR TAMIL QA: HARNESSING THE WISDOM OF NANNŪL FOR ACCURATE ANSWERS

NIVEDITHA S[1] , PAAVAI ANAND G[2]

[1]Department of Computer Science and Engineering, Faculty of Engineering and Technology,

SRM Institute of Science and Technology, Vadapalani, Chennai, India.

[2]Department of Computer Science and Engineering, Faculty of Engineering and Technology,

SRM Institute of Science and Technology, Vadapalani, Chennai, India.

E-mail: [1]nivedits@srmist.edu.in, [2]paavaiag@srmist.edu.in

ABSTRACT

Morpheme Mastery represents a groundbreaking advancement in Tamil Question Answering, introducing a pioneering answer alignment technique that leverages Finite State Transducer (FST) technology to expertly handle complex morphological structures, particularly nouns. This research contributes to the advancement of IT in natural language processing by developing a novel approach that ensures precise analysis and interpretation, enabling efficient and accurate processing. The significance of this research lies in its potential to revolutionize computational linguistics and Tamil Question Answering, setting a new benchmark for accuracy and efficiency. By incorporating novel steps such as stop words removal, Numeric Information Extraction (NIE), and Temporal Entity Recognition (TER), Morpheme Mastery addresses a critical gap in existing question answering systems. Our deliberate choice of a rule-based approach, rather than machine learning alternatives, guarantees consistently high accuracy, a critical prerequisite for building a robust computational language framework. Empirical results demonstrate the efficacy of our approach, yielding an impressive 11.87% improvement in CHAII and 9.58% in SQuAD dataset efficiency. This significant enhancement underscores Morpheme Mastery's potential to transform the landscape of Tamil Question Answering and computational linguistics, enabling more accurate and efficient human-machine interaction.

Keywords: *Answer Alignment, Tamil, Question Answering, Morpheme Analysis, Nannūl*

1. INTRODUCTION

Question-answering (QA) systems have evolved significantly, from simple information retrieval to complex conversational AI. Comprehensive surveys highlight the progression from single-turn to multi-turn QA, emphasizing the importance of understanding context and engaging in dialogue [1]. QA systems typically comprise three core components: question classification, information retrieval, and answer extraction [2]. These systems employ various representation levels, from basic keyword searches to complex logical queries, integrating natural language processing techniques such as part-of-speech tagging and semantic role labelling [3]. Community Question Answering (CQA) platforms have emerged as popular Web 2.0 applications, harnessing collective intelligence to address complex, subjective queries [4]. Recent advancements in pre-trained language models and the availability of large-scale, multi-turn QA datasets have further propelled the field, paving the way for more sophisticated conversational AI systems [1].

Tamil question-answering (QA) systems face significant challenges due to the language's inherent complexity and limited resources. Tamil, as an agglutinative language, forms words by combining smaller units like roots, prefixes, and suffixes, allowing a single word to express intricate meanings through markers for tense, gender, mood, case, and number. This rich linguistic structure, while enabling nuanced expression, results in a vast number of possible word forms, complicating computational analysis. Verbs in Tamil, for example, exhibit extensive inflection, encompassing tense, mood, voice, and politeness, making accurate parsing and analysis a demanding task for NLP models. These structural intricacies are compounded by the scarcity of standardized resources such as annotated corpora, tokenizers, and morphological analyzers. [8] focused on Tamil question classification using a Conditional Random Field model based on morpheme features, which discriminates the position of expected answer type information. The lack of high-quality Tamil QA datasets further exacerbates the issue, forcing researchers to explore alternatives like zero-shot transfer, multilingual models, and dataset creation through translation [7]. Despite these obstacles, advancements have been made in Tamil NLP. For instance, domain-specific QA [6] systems have been developed using rule-based and machine-learning techniques for Tamil history texts, achieving promising results with models like XLM-RoBERTa on extractive QA tasks [5]. However, challenges remain, such as generating grammatically correct and meaningful Tamil questions, with only 62.22% meeting the required standards [6]. Named Entity Recognition (NER), a critical component of QA systems, adds another layer of complexity when dealing with diverse web data sources.[10]

Morphological complexity poses significant challenges for question-answering (QA) systems, particularly in languages with rich morphology. These complexities can exacerbate the lexical gap between questions and answers [11]. To address this, researchers have explored various morphological resources and techniques. German, [12] developed a QA system utilizing morphological transformations to improve recall. In Arabic, the rich morphology necessitates preprocessing for effective statistical modelling [13]. French, [14] conducted a detailed analysis of constructional morphology in QA, creating an annotated corpus to evaluate morphological resources and tools. [11] specifically examined deverbal agent nouns in French QA.

Morphology, a fundamental branch of linguistics, examines words' structure, formation, and internal construction, focusing on components like roots, stems, prefixes, suffixes, and inflectional patterns to convey grammatical meanings such as tense, number, gender, and case. This understanding is crucial in linguistics and Natural Language Processing (NLP), enabling deeper word-level analysis and processing. Tamil, an agglutinative Dravidian language with rich morphological features, exemplifies the challenges and opportunities in this field. Despite limited extensive study, researchers have made significant strides in Tamil morphological analysis using various methods. Machine learning techniques have shown promise, with [15] achieving 95.65% accuracy through sequence labelling and kernel methods, and [16] attaining 98.73% accuracy using support vector machines with linguistic features. Rule-based approaches have also proven effective, as demonstrated by [18] in analyzing classical Tamil texts. In the gynecology domain, [17] found a paradigm-based finite state model to be the most effective, achieving an accuracy of 0.96. However, this rule-based approach to morphological analysis can be limited by its dependence on a dictionary and sequential rule application, making it vulnerable to errors if a single rule fails. While the Machine Learning (ML) approach mitigates the vulnerability of cascaded rule failure inherent in rule-based systems, it introduces a new dependency on corpora. However, relying solely on corpora poses a significant limitation, as they may not encompass an exhaustive list of Tamil words, rendering them incomplete and potentially inaccurate for comprehensive language analysis.

Finite-state models have proven valuable in natural language processing and question-answering systems. They have been applied to question classification, achieving high accuracy using trained finite state machines (FSMs) with simple learning approaches [19]. In parsing, finite-state approximations of wide-coverage parsers, implemented as hidden Markov models, have been used to improve efficiency [20]. Finite-state devices, including automata, graphs, and transducers, have applications in dictionary encoding, text processing, and speech recognition [21]. In complex question answering over knowledge bases, a state transition-based approach has been proposed to translate natural language questions into semantic query graphs using primitive operations and learning-based transitions [22]. These diverse applications demonstrate the versatility and effectiveness of finite-state models in handling various natural language processing tasks, particularly in question-answering systems.

The proposed Tamil QA system represents a significant contribution to the field of Information Technology, particularly in the development of language technologies for low-resource languages. By leveraging advanced natural language processing techniques and finite-state transducer technology, this research aims to push the boundaries of language accessibility and usability, ultimately enhancing the effectiveness of information dissemination and retrieval for millions of Tamil speakers worldwide.

2. METHODOLOGY

Our system employs a multi-layered validation framework, comprising four distinct phases: preprocessing and three tiers of scrutinization. The preprocessing phase initiates with the removal of stop words, followed by syntactic alignment verification to ensure semantic consistency. This is followed by three scrutinization steps: Temporal Entity Recognition to identify and contextualize temporal entities, Numerical Information Extraction to extract and process numerical data, and Morphological structures, ensuring linguistic Alignment to accuracy and precision.

As shown in Figure 1, the crucial step in our question answering framework involves categorizing questions based on the presence of specific question words. This categorization yields three primary types: Temporal, Numerical, and Non-Temporal Non-Numerical Questions. Temporal questions, characterized by the presence of words like "*எப்போது*" [When] (eppōtu) or "*எப்பொழுது*" [When] (eppol̲utu) anticipate a date as the response.

Consequently, both the retrieved answer and the dataset answer are aligned in the standardized format "Month DD, YYYY." Conversely, Numerical questions, marked by words such as "**எத்தனை**" (ettan̲ai) [How much] or "**எவ்வளவு**" (evvaḷavu) [How much] expect a value as the response, prompting the alignment of answers in the "Number Units" format. All remaining questions fall under the " Non-Temporal Non-Numerical Questions" category, for which morphological alignment is employed to check for morpheme equivalence.

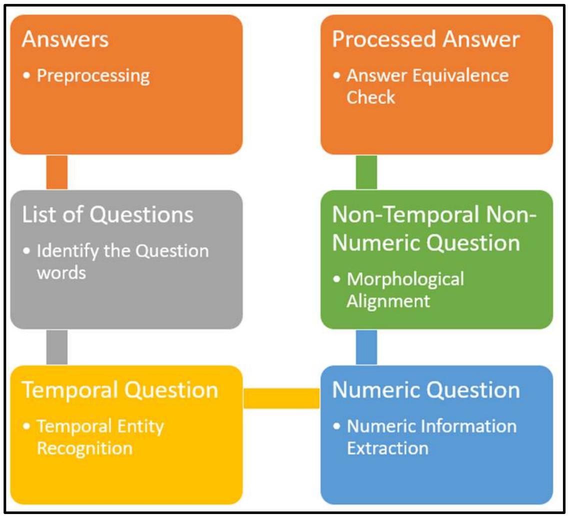
Answers
• Preprocessing
Processed Answer
• Answer Equivalence Check
List of Questions
• Identify the Question words
Non-Temporal Non-Numeric Question
• Morphological Alignment
Temporal Question
• Temporal Entity Recognition
Numeric Question
• Numeric Information Extraction

Figure 1. Architecture diagram

2.1 Preprocessing

The preprocessing stage involves two pivotal operations: Stop Word Removal and Syntactic Alignment Verification. The following gives the detailed explanation about the same.

2.1.1. Stop word removal

The elimination of inconsequential words, known as stop words, as shown in Table 1. is a crucial step in optimizing text processing. These ubiquitous words, such as Tamil equivalents of "the", "and", “some”, “about” etc., do not contribute substantially to the text's meaning. By removing them, processing efficiency is significantly enhanced. A comprehensive analysis revealed that Tamil, like English, possesses its own set of stop words that serve no functional purpose, and thus, are systematically eliminated.

2.1.2. Syntactic alignment verification

The process commences with a thorough examination of the answer to ascertain whether it comprises exclusively numerical values. However, a comprehensive analysis of the SQuAD dataset uncovered 16 instances where numerical values were erroneously distinguished due to discrepancies in spacing and extraneous zeros. To rectify this issue, we devised a robust normalization procedure to ensure uniform numerical value representation, thereby enhancing accuracy. The following table illustrates a few exemplary instances of this phenomenon. To verify the equivalence of numerical values, a simple yet efficacious two-step approach is employed:

Space Normalization: Extraneous spaces are removed to ensure consistent formatting.

Value Normalization: The numerical value is divided by 100 to facilitate accurate comparison.

Although these steps may seem elementary, they prove to be highly effective in determining equivalence between numerical values. A sample is given in Table 2.

A series of Finite State Transducers (FSTs) are utilized to accomplish tasks such as question categorization, Temporal Entity Recognition, and Numerical Information Extraction. To address intricate details, Algorithm 1 is employed to illustrate morpheme alignment.

2.2. Question Categorization

To facilitate accurate answer alignment, the initial step involves identifying the question type. This is crucial, as the question type dictates the anticipated pattern for the answers. Consequently, question type analysis is performed based on the presence of specific question words. As illustrated in Figure 2, the presence of question words "*எப்போது*"

[When] (eppōtu) or "எப்பொழுது" [When] (eppolutu) indicates that the answer is expected to be a date.

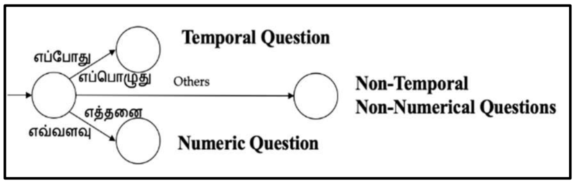

Figure 2. Finite State Transducer for Question categorization

2.2.1. Temporal question type

In temporal question case, the date format is expected to conform to the pattern depicted in Fig 3: Month DD, YYYY. This format comprises the full month name, a two-digit date, and a four-digit year representation.

Notably, the month can be expressed in either English or Tamil, with the 12 possible months listed in Fig 4. The Finite State Transducers (FSTs) presented in Figures 5 and 6 demonstrate how two-digit dates and four-digit years are recognized, respectively. While the expected answer format typically consists of a numeric value, space, and unit name, variations may exist in the dataset. In such cases, as shown in Table 3 the retrieved answer is predicted based on the available elements, resulting in alternative patterns, such as: YYYY (year only), Month YYYY (month and year), Month DD (month and day). These variations ensure flexibility in answer prediction, accommodating diverse formats present in the dataset

Table 1 Stop word removal

Question	Answer in dataset	Answer
தென் கொரிய KPA ஐத் தூண்டவில்லை என்ற கூற்றை யார் கேள்வி எழுப்பினர்? Teṉ koriya KPA ait tūṇṭavillai eṉṟa kūṟṟai yār kēḷvi eḻuppiṉar? Who questioned the claim that the South Koreans did not provoke the KPA?	அறிஞர்கள் Aṟiñarkaḷ Scholars	சில அறிஞர்கள் Cila aṟiñarkaḷ Some scholars
எந்த கட்டத்தில் புல்வெளியில் வோர்ட் பீர் ஆனது? Enta kaṭṭattil pulveḷiyil vōrṭ pīr āṉatu? At what point did wort turn into beer in the meadow?	நொதித்தல் notittal fermentation	நொதித்தல் போது notittal pōtu during fermentation
சுவிட்சர்லாந்தில் எத்தனை பேர் வேலை செய்கிறார்கள்? Cuviṭcarlāntil ettaṉai pēr vēlai ceykiṟārkaḷ? How many people work in Switzerland?	சுமார் 3.8 மில்லியன் cumār 3.8 Milliyaṉ about 3.8 million	3. 8 மில்லியன் 3. 8 Milliyaṉ 3. 8 million
பூகம்பத்தின் காரணமாக, எத்தனை பேர் வீட்டுவசதி இல்லை? Pūkampattiṉ kāraṇamāka, ettaṉai pēr vīṭṭuvacati illai? How many people are homeless because of the earthquake?	குறைந்தது 5 மில்லியன் kuṟaintatu 5 milliyaṉ at least 5 million	5 மில்லியன் 5 milliyaṉ 5 million
எவ்வளவு காலமாக ஃபயர்ஸ்டோன் டயர் மற்றும் ரப்பர் நிறுவனம் லைபீரியாவில் ஒரு ரப்பர் தோட்டத்தை இயங்கின? Evvaḷavu kālamāka ḥpayarsṭōṉ ṭayar maṟṟum rappar niṟuvaṉam laipīriyāvil oru rappar tōṭṭattai iyaṅkiṉa? How long did Firestone Tire and Rubber Company operate a rubber plantation in Liberia?	1926 முதல் 1926 mutal since 1926	1926

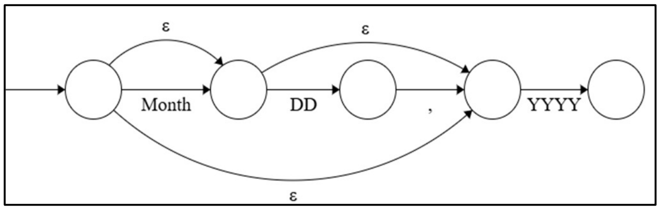

Figure 3. Temporal Entity Recognition
Table 2. Space and Numerical Normalization

Question	Answer in dataset	Retrieved Answer
2013-4 ஆம் ஆண்டில், நியூயார்க் மாநில வரி வருவாயில் என்ன சதவீதம் வோல் ஸ்ட்ரீட்டில் பத்திரங்கள் வணிகத்திலிருந்து வந்தது? 2013-4 Ām āṇṭil, niyūyārk mānila vari varuvāyil eṉṉa catavītam vōl sṭrīṭṭil pattiraṅkaḷ vaṇikattiliruntu vantatu? In 2013-4, what percentage of New York State tax revenue came from securities trading on Wall Street?	19%	19%
நைஜீரியாவின் ஆண் மக்கள் எவ்வளவு படிக்க முடியும்? Naijīriyāviṉ āṇ makkaḷ evvaḷavu paṭikka muṭiyum? How literate is the male population of Nigeria?	75.70%	75. 7 %
மக்கள் தொகையில் என்ன சதவீதம் ஹிஸ்பானிக் என அடையாளம் காட்டுகிறது? Makkaḷ tokaiyil eṉṉa catavītam hispāṉik eṉa aṭaiyāḷam kāṭṭukiṟatu? What percentage of the population identifies as Hispanic?	28.60%	28. 6 %
எவ்வளவு பிரேசிலியாவின் மொத்த உள்நாட்டு உற்பத்தியில் பொது நிர்வாகத்தில் இருந்து வருகிறது? Evvaḷavu pirēciliyāviṉ motta uḷnāṭṭu uṟpattiyil potu nirvākattil iruntu varukiṟatu? How much of Brasilia's GDP comes from public administration?	54.80%	54. 8 %

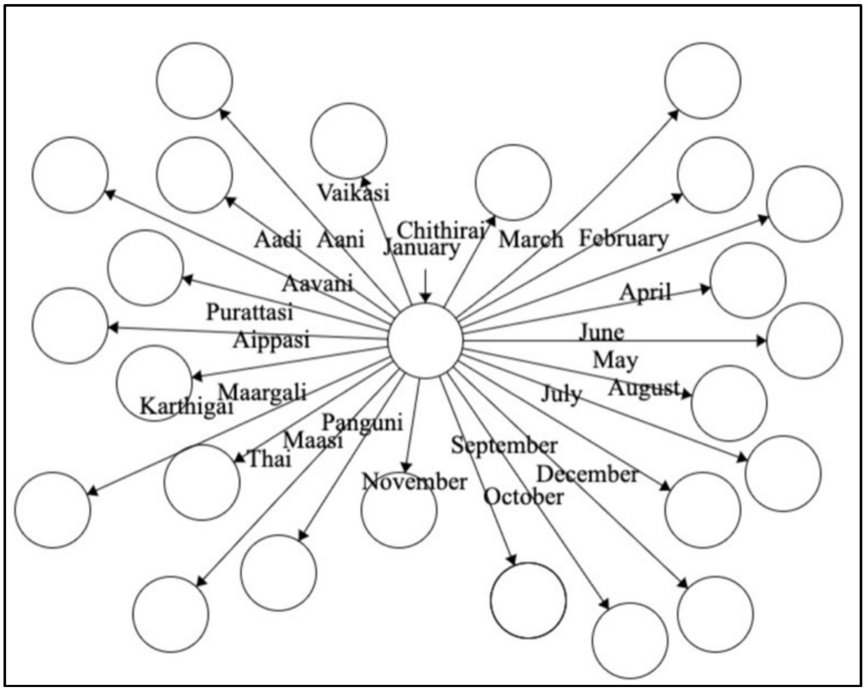

Figure 4. Finite State Transducer for Month
Table 3 Temporal Entity Recognition

Question	Answer in dataset	Retrieved Answer

இயந்திரப் பொறியாளர் ஜேம்ஸ் வாட் எப்போது பிறந்தார்? Iyantirap poṟiyāḷar jēms vāṭ eppōtu piṟantār? When was mechanical engineer James Watt born?	1736 ஆம் ஆண்டு ஜனவரி மாதம் 19 1736 Ām āṇṭu jaṉavari mātam 19 January 19, 1736	ஜனவரி 19, 1736 Jaṉavari 19, 1736
இந்து செய்தித்தாள் நிறுவனம் எப்போது நிறுவப்பட்டது? Intu ceytittāḷ niṟuvaṉam eppōtu niṟuvappaṭṭatu? When was The Hindu Newspaper Company founded?	1878	செப்டம்பர் 20, 1878 ceptampar 20, 1878
திமுக அரசியல் கட்சித் தலைவர் கருணாநிதி எப்போது பிறந்தார்? Timuka araciyal kaṭcit talaivar karuṇāniti eppōtu piṟantār? When was DMK political party leader Karunanidhi born?	1924 சூன் 3 1924 cūṉ 3	சூன் 3, 1924 cūṉ 3, 1924

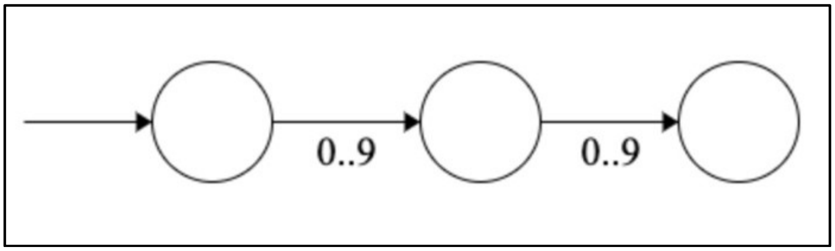

Figure 5. Finite State Transducer for MM

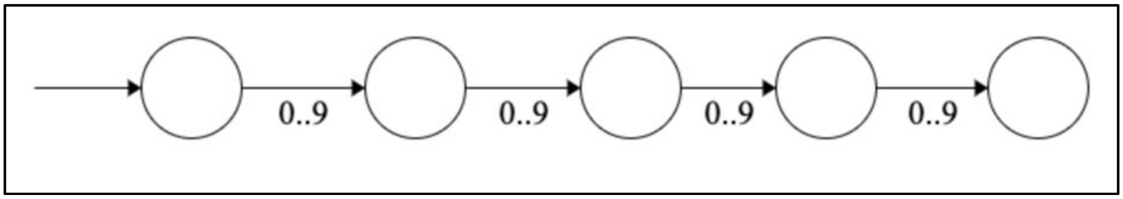

Figure 6. Finite State Transducer for YYYY

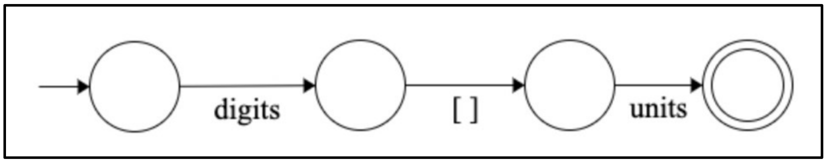

Figure 7. Numeric Information Extraction

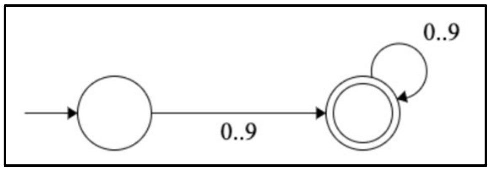

Figure 8. Finite State Transducer for digits

Table 4 Numeric Information Extraction

Question	Answer in dataset	Retrieved Answer
இன்று எத்தனை பேர் ரோமன் கத்தோலிக்கர்கள்? Inṟu ettaṉai pēr rōmaṉ kattōlikkarkaḷ? How many people are Roman Catholics today?	1.17 பில்லியன் 1.17 Pilliyaṉ 1.17 billion	1. 17 பில்லியன் 1.17 Pilliyaṉ 1.17 billion
இஸ்ரேலின் இறையாண்மை எவ்வளவு பெரியது? Isrēliṉ iṟaiyāṇmai evvaḷavu periyatu? How big is the sovereignty of Israel?	20,770 சதுர கிலோமீட்டர் 20,770 catura kilōmīṭṭar 20,770 square kilometers	20, 770 சதுர கிலோமீட்டர் 20,770 catura kilōmīṭṭar 20,770 square kilometers
பசிபிக் பெருங்கடல்லின் அதிகபட்ச ஆழம் எவ்வளவு? Pacipik peruṅkaṭalliṉ atikapaṭca āḻam evvaḷavu? What is the maximum depth of the Pacific Ocean?	10911	10, 911 மீட்டர் 10, 911 Mīṭṭar 10,911 meters

Table 5. Noun Stem Category Classification

Category	Suffix 1	Suffix 2	Suffix 3	Suffix 4
Nominative	No suffix			
Accusative	ஐ (ai)			
Instrumental and Sociative cases	ஆல் (āl)	ஆன் (āṉ)	ஒடு (oṭu)	ஓடு (ōṭu)
Dative form	கு (ku)	ஆக (āka)		
Ablative case	இல் (il)	இன் (iṉ)	இருந்து (iruntu)	
Genitive case	அது (atu)	ஆது (ātu)	உடைய (uṭaiya)	
Locative case	கண் (kaṇ)	இல் (il)	உள் (uḷ)	இடம் (iṭam)
Vocative form	ஏ(ē)	ஓ(ō)		

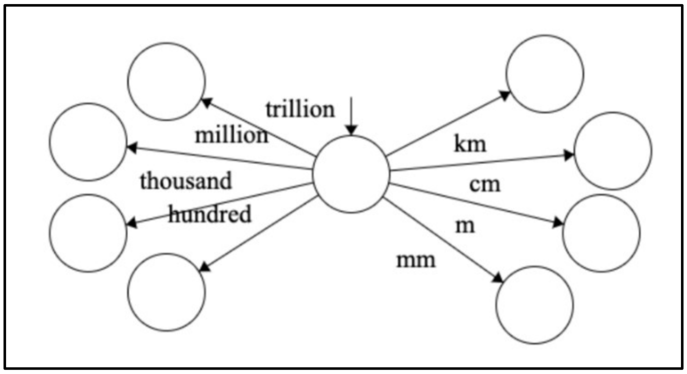

Figure 9. Finite State Transducer for Units

Table 6. Locative Noun

Question	Answer in dataset	Retrieved Answer
பீர் ஒரு பாதுகாப்பற்ற என்ன மூலப்பொருள் செயல்படுகிறது அமிலத்தன்மை? Pīr oru pātukāppaṟṟa en̲n̲a mūlapporuḷ ceyalpaṭukiṟatu amilattan̲mai? What ingredient in beer acts as an unsafe acidity?	ஹாப்ஸ்	ஹாப்ஸின்
Patrick Modiano எந்த நகரத்தில் வாழ்கிறது? Patrick Modiano enta nakarattil vāl̲kiṟatu? What city does Patrick Modiano live in?	பாரிஸ்	பாரிஸில்
தென் கொரியாவில் நடைபெற்ற டார்ச் ரிலே நிகழ்வு எங்கே? Ten̲ koriyāvil naṭaipeṟṟa ṭārc rilē nikal̲vu eṅkē? Where was the torch relay held in South Korea?	சியோல்	சியோலில்

Table 7. Dative Noun

Question	Answer in dataset	Retrieved Answer
கூடுதல் கடன் வழங்குவதற்கு போதுமான மூலதனத்தை உருவாக்க எத்தனை ஆண்டுகள் வலுவான இலாபம் பெறும்? Kūṭutal kaṭan̲ val̲aṅkuvataṟku pōtumān̲a mūlatan̲attai uruvākka ettan̲ai āṇṭukaḷ valuvān̲a ilāpam peṟum? How many years will it take to generate strong profits to build up enough capital to provide additional loans?	பல ஆண்டுகள் Pala āṇṭukaḷ Many years	பல ஆண்டுகளுக்கு Pala āṇṭukaḷukku For many years
அன்வார் எல் சதாத் ஒரு பயணம் எங்கு சென்றார்? An̲vār el catāt oru payaṇam eṅku cen̲ṟār? Where did Anwar El Sadat go on a trip?	இஸ்ரேல் Isrēl Israel	இஸ்ரேலுக்கு isrēlukku to Israel

Table 8. Other Nouns

Question	Answer in dataset	Retrieved Answer
1975 ஆம் ஆண்டில் முதல் தடவையாக ராணி என்ன நாட்டில் விளையாடினார்? 1975 Ām āṇṭil mutal taṭavaiyāka rāṇi en̲n̲a nāṭṭil viḷaiyāṭin̲ār? Which country did Queen play for the first time in 1975?	கனடா Kan̲aṭā Canada	கனடாவில் Kan̲aṭāvil in Canada
இரண்டாம் உலகப் போரின் முடிவில் என்ன நாடு பிரிக்கப்பட்டது? Iraṇṭām ulakap pōrin̲ muṭivil en̲n̲a nāṭu pirikkappaṭṭatu? Which country was partitioned at the end of World War II?	கொரியா Koriyā Korea	கொரியாவின் koriyāvin̲ Korea
டெக்கான் பொறிகளை எங்கே? Ṭekkān̲ poṟikaḷai eṅkē? Where are the Deccan Traps?	இந்தியா intiyā India	இந்தியாவில் Intiyāvil In India

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

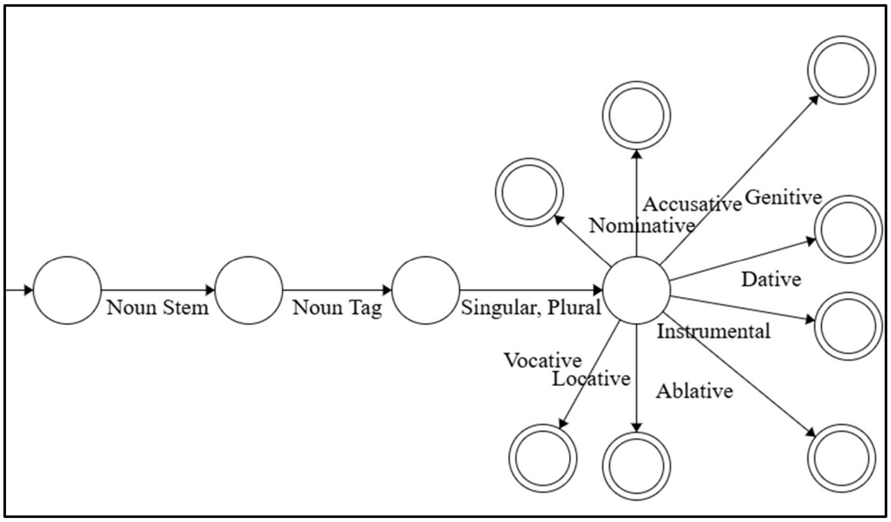

Figure 10. Noun Transducer [24]

Table 9. Results

Category	SQuAD	CHAII
Numerical	16	19
Locative	3	1
Dative	2	1
Other Noun cases	40	1
Unsolved types	11	0

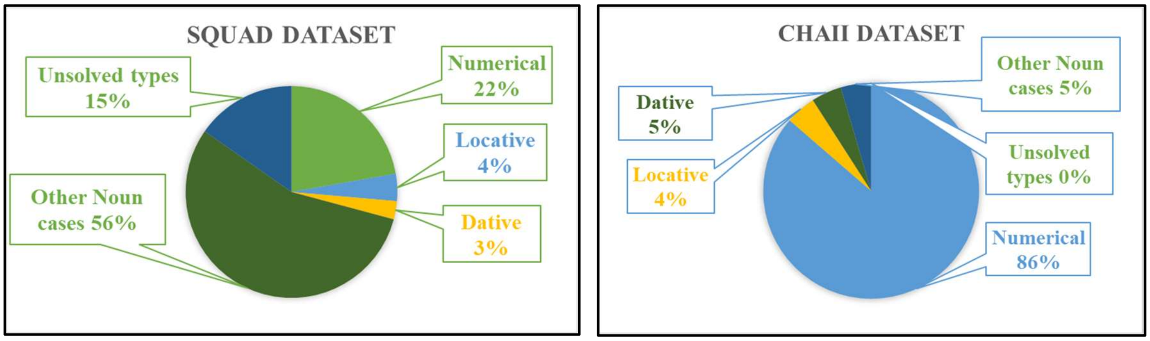

Figure 11a. Performance of SQuAD Dataset and 11b. CHAII Dataset

2.2.2. Numeric question type

Figure 2 reveals that the presence of question words "எத்தனை" (ettaṉai) [How much] or "எவ்வளவு" (evvaḷavu) [How much] signals that the answer is expected to be a quantitative value. Consequently, the anticipated answer format consists of a numeric value, followed by a space [ ], and then the unit name, as illustrated in Figure 7 and in Table 4.

The Finite State Transducer (FST) for recognizing numeric values is presented in Figure 8. Additionally, Figure 9 showcases various possible unit variations, highlighting the flexibility in unit representations.

2.2.3 Non-Temporal Non-Numerical question type

Tamil's morphological structure is notably intricate [9], as it involves inflectional markings for person, gender, and number and auxiliary combinations that convey aspect, mood, causation, and attitude in verbs.
Nouns, meanwhile, undergo inflection with suffixes indicating plurality, obliqueness, case, postpositions, and clitics. To analyze languages with such complex inflectional systems accurately, it is essential to identify the root and morphemes of each word.

Tamil grammar, Nannūl [23] comprises eight distinct noun suffix classes, each encompassing multiple suffixes. These classes govern the various grammatical cases in Tamil, including:
Type 1: Nominative form (no suffix)
Type 2: Accusative form
Type 3: Instrumental and Sociative cases
Type 4: Dative form
Type 5: Ablative case
Type 6: Genitive case
Type 7: Locative case
Type 8: Vocative form (used for addressing or calling a person)

As shown in Table 5., these eight classes collectively contain approximately 20 suffixes, which, when applied to both singular and plural forms, yield around 40 distinct forms. Tamil nouns can be categorized into seven stem categories based on morphological behavior.
The finite state model for the noun classification is shown in Fig 10.

This category encompasses a diverse range of question types that do not conform to specific patterns. Despite the absence of a predefined structure, a meticulous examination of morphemes is conducted to facilitate accurate answer alignment, ensuring that subtle linguistic cues are carefully considered.

A detailed explanation of this algorithmic process is provided in the subsequent section, offering a comprehensive understanding of the system's operational framework.

```
Algorithm Morphological Alignment (S):
    Remove all special characters from S
    Determine grammatical case C of S
    if C == locative:
        Remove last suffix
        Change second last uyir_mei to mei
    elif C == dative:
        Remove last 2 suffixes
        Change third last uyir_mei to mei
    elif C in {accusative, ablative, genitive, vocative}:
        Remove last suffix
    Return the processed string
```

Algorithm 1. Morphological Alignment

A 3-stage normalization protocol is employed to standardize Morphological alignment.

Stage 1: Locative case
For locative case, the terminal suffix is excised. Subsequently, the penultimate uyir-mei letter undergoes transformation, wherein the uyir component is detached, yielding the corresponding mei letter. The instances are listed in the Table 6
For instance, consider the retrieved answer, ஹாப்ஸின்:
1. Terminal suffix removal: ஹாப்ஸி
2. Penultimate uyir-mei letter extraction: ஸி
3. Uyir component detachment: ஸ் (mei letter)
4. Normalized output: ஹாப்ஸ்

Stage 2: Dative case
For dative case, last 2 suffixes are removed. Subsequently, the third suffix undergoes transformation, wherein the uyir component is detached, yielding the corresponding mei letter. The instances are listed in the Table 7

1. Dual suffix excision: The terminal and penultimate suffixes are removed.
2. Uyir-mei transformation: The third-last uyir-mei letter undergoes conversion, detaching the uyir component to yield the corresponding mei letter.

For instance, consider the retrieved answer, இஸ்ரேலுக்கு:

1. Dual suffix removal: இஸ்ரேலு
2. Uyir-mei transformation: இஸ்ரேல்

Stage 3: Accusative, Ablative, Genitive, Vocative

A unified suffix removal strategy is employed for non-locative and non-dative cases, wherein excision of the terminal two suffixes consistently yields the root word. This streamlined approach facilitates efficient morphological normalization across various case categories. Few of the instances are listed in the Table 8

For instance, consider the retrieved answer கனடாவில்:

After removing the last 2 suffixes, கனடா

3. RESULTS AND DISCUSSION

We have run our algorithm on existing Tamil QA datasets, namely SQuAD [25] and CHAII [28] datasets comprising of 1201 and 368 question-answer pairs respectively, serves as a benchmark for comprehensive Question Answering (QA) systems. In this study, we leveraged the MURIL BERT [27] model as our QA system and employed the Jaccard measure to assess its performance.

As given in Table 9, our analysis revealed that approximately 72 questions in SQuAD and 22 questions in CHAII dataset posed challenges related to understanding morphologically equivalent words. To address this, we employed our Morpheme mastery model, which successfully resolved the issues in these cases. Notably, 16 questions in SQuAD and 19 questions in CHAII dataset involved numerical equivalence issues, which were effectively addressed using our Numerical normalization technique.

A deeper examination of the challenges revealed that understanding morphological equivalence was the primary concern. To tackle this, we categorized the instances based on noun classification. Our results showed that all locative and dative cases (5 instances in SQuAD and 2 instances in CHAII) were successfully resolved. Furthermore, our approach effectively addressed the remaining 40 instances in SQuAD and 1question in CHAII involving other noun cases. An improvement in efficiency of 84.72% in SQuAD and 100% in CHAII was achieved.

Figure 11a illustrates the distribution of 72 Question-Answering (QA) pairs in SQuAD dataset, revealing a categorical breakdown of challenges encountered. Notably, 22% (16 instances) of QA pairs involved numerical answers, which were successfully resolved through numerical normalization techniques. Locative and Dative-related queries accounted for 4% (3 instances) and 3% (2 instances) of the total QA pairs, respectively, corresponding to locative and dative cases.

These cases were effectively addressed through morphological analysis and suffix removal strategies. Furthermore, a substantial proportion (56%; 40 instances) of QA pairs pertained to other noun cases, which were also successfully resolved using a combination of morphological and syntactic analysis techniques. Collectively, these results indicate that 85% (61 instances) of QA pairs were successfully resolved, with only 15% (11 instances) remaining as outliers that could not be categorized under the predefined noun cases.

Figure 11b illustrates the distribution of 22 Question-Answering (QA) pairs in the CHAII dataset, revealing a categorical breakdown of challenges encountered. A significant majority, 86%, required numerical answers, which were resolved using numerical normalization techniques. Locative and Dative-related queries accounted for 4% and 5%, respectively, and were addressed through morphological analysis and suffix removal. Another 5% pertained to other noun cases, resolved using morphological and syntactic analysis. The model successfully resolved 100% of QA pairs, with 0% remaining as outliers.

The efficacy of our algorithm is vividly illustrated in the graph depicted in Fig 12, which provides a comprehensive breakdown of the total number of cases identified across each dataset. Notably, the graph also delineates the distribution of instances into distinct categories, including numerical, locative, dative, and other noun cases, as well as outliers. This granular visualization facilitates a deeper understanding of the algorithm's performance and its ability to accurately categorize and identify various cases within the datasets.

As shown in the Table 10, the CHAII dataset shows the performance of a model on three types of questions: Temporal Data Based, Numerical Data Based, and Non-Temporal Non-Numerical Based.

The overall performance improves by 11.87% after Morpheme Mastery which includes Preprocessing, Temporal Entity Recognition, Numerical Information Extraction and Morphological alignment.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

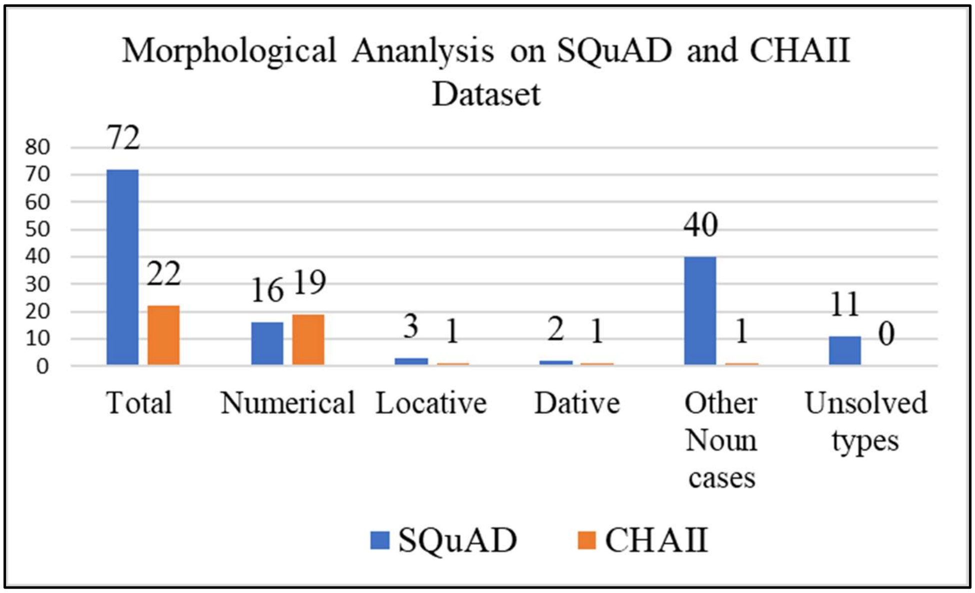

Figure 12. Morphological Analysis on CHAII and SQuAD Dataset

Table 10. Morpheme Mastery on CHAII dataset

CHAII Dataset	Total	No. of Correct Answers	Exact Match	After Preprocessing	Exact Match	After Morpheme Mastery	Exact Match
Temporal Data Based Questions	70	41	58.57	43	61.43	55	78.57
Numerical Data Based Questions	46	24	52.17	28	60.87	29	63.04
Non-Temporal Non-Numerical Based Data	252	137	54.37	151	59.92	158	62.70
TOTAL	368	202		222		242	

Table 11 Morpheme Mastery on SQuAD dataset

SQuAD Dataset	Total	No. of Correct Answers	Exact Match	After Preprocessing	Exact Match	After Morpheme Mastery	Exact Match
Temporal Data Based Questions	19	7	36.84	10	52.63	11	57.89
Numerical Data Based Questions	135	57	42.22	89	65.92	90	66.67

Non-Temporal Non-Numerical Based Data	1047	376	35.91	414	39.54	454	44.12
TOTAL	1201	440		513		555	

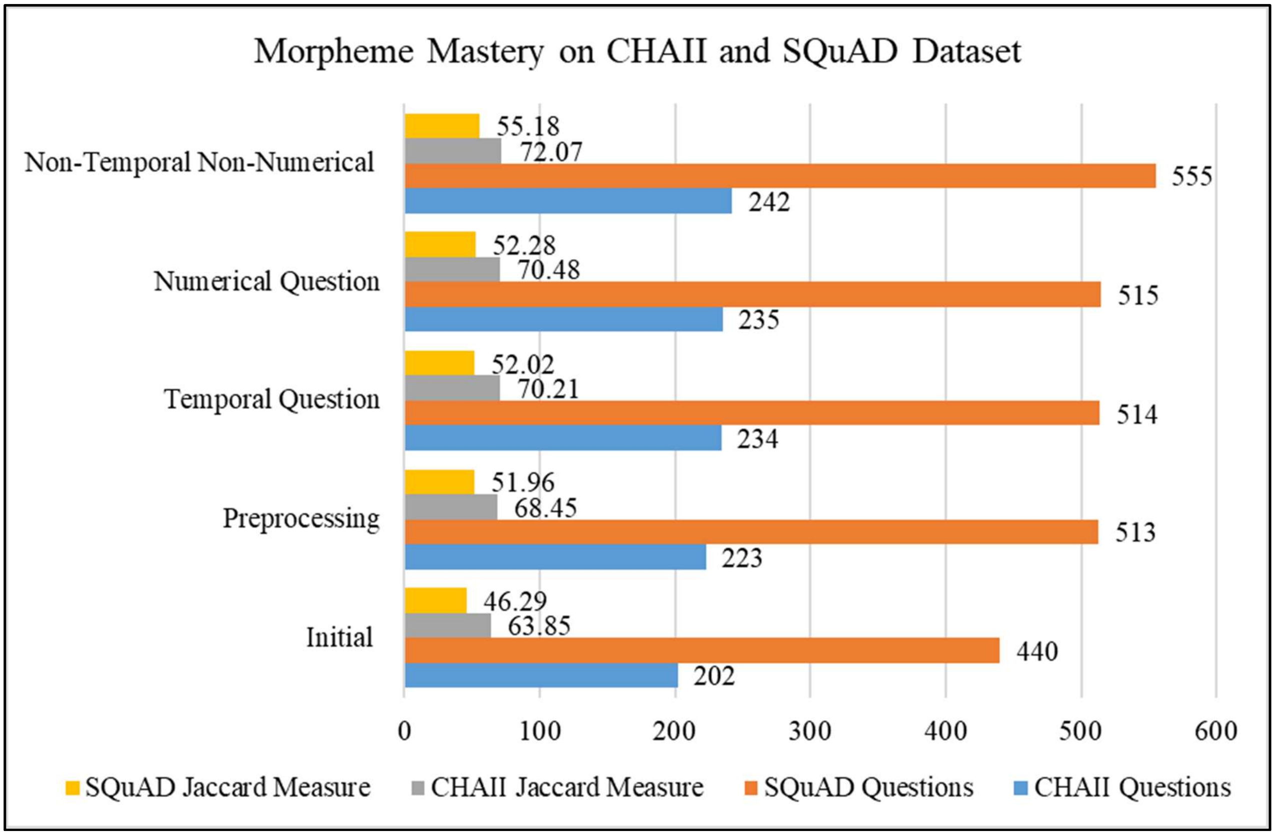

Figure 13. Comparative Study

1. **Temporal Questions**: The question for which the expected answers were dates with different formats, when handled properly, the exact match rate increases from 58.57% to 78.57%, showing a significant improvement of 20%, thereby showcasing enhanced proficiency in handling diverse date formats.

2. **Numerical Questions**: The questions requiring answers in the format of numerical value accompanied by its corresponding units, witnessed a notable improvement in exact match rate, rising from 52.17% to 63.04%, showcasing enhanced accuracy in handling numerical values with units

3. **Non-Temporal Non-Numerical Questions**: The questions heavily influenced by morpheme analysis, which comprise a substantial portion of the dataset, demonstrated considerable improvement. The exact match rate surged from 54.37% to 62.70%, underscoring enhanced proficiency in morpheme analysis and its significant impact on question answering accuracy.

Overall performance over CHAII dataset is 54.89% while after Morpheme mastery it has increased to 65.76%. As shown in Table 11, the SQuAD dataset evaluates a model's performance across three question types. Following Morpheme Mastery - encompassing preprocessing, temporal entity recognition, numerical info extraction, and morphological alignment - overall performance jumps 9.58%.

1. **Temporal Questions:** Proper handling of diverse date formats boosts exact match rates by 21.05%, from 36.84% to 57.89%.
2. **Numerical Questions:** Enhanced accuracy in handling numerical values with units increases exact match rates by 24.45%, from 42.22% to 66.67%.
3. **Non-Temporal Non-Numerical Questions:** Morpheme analysis proficiency improves exact match rates by 8.21%, from 35.91% to 44.12%.
Overall performance over SQuAD dataset is 36.63% while after Morpheme mastery it has increased to 46.21%

A comparative analysis of CHAII and SQuAD datasets reveals distinct characteristics as shown in Fig 13. Notably, the CHAII dataset comprises a predominant number of temporal questions, necessitating robust handling by Morpheme Mastery. In contrast, the SQuAD dataset presents a substantial number of questions that require precise Morpheme alignment, underscoring the importance of this technique in achieving accuracy.

4. CRITICAL EVALUATION OF FINDINGS

Upon completing a thorough investigation of the previously discussed domains, our research identified several additional issues that could be resolved. A detailed exposition of our discoveries is presented below.

4.1. Critique of Results

Initially, our investigation focused on morphological analysis based on noun stem classification, as delineated in Table 5. However, subsequent in-depth examination of our dataset revealed that morphological applications transcended the classifications presented in Table 5. Notably, additional suffix variants were identified, a subset of which are presented in Table 12. This finding underscored the complexity of the Tamil language, which supports a vast array of suffixes, not all of which could be exhaustively addressed within the scope of this study.

Table 12. Data Mavericks

Question	Answer in dataset	Retrieved Answer
ஜெர்மனியில் மிகப்பெரிய மதம் எது? *Jermaṉiyil mikapperiya matam etu?* *What is the largest religion in Germany?*	கிறித்தவத்தை *Kiṟittavattai* *Christianity*	கிறித்தவ *Kiṟittava* *Christianity*
என்ன பெரிய பூனை நாய்கள் தாக்க ஒரு போக்கு உள்ளது? *Eṉṉa periya pūṉai nāykaḷ tākka oru pōkku uḷḷatu?* *What big cat has a tendency to attack dogs?*	சிறுத்தை *Ciṟuttai* *Leopard*	சிறுத்தைகள் சிறுத்தைகள் *Leopards*

Similarly, our analysis revealed that Named Entity Recognition (NER) posed significant challenges. Specifically, instances in the dataset demonstrated that identical names were addressed differently, depending on the context, thereby affecting the model's interpretation and accuracy. Few instances are shown in the Table 13.

Additionally, other areas of concern that warrant attention include:

1. Stop word removal: Ensuring the effective elimination of common words that do not add significant value to the text analysis.

2. Foreign language handling: Developing strategies to accommodate and process text data in languages other than the primary language of focus.

Table 13. Naming Conundrums

Question	Answer in dataset	Retrieved Answer
கல்லணை கட்டியது யார்? *Kallaṇai kaṭṭiyatu yār?* *Who built the kallanI?*	கரிகாலன் *Karikālaṉ* *Karikalan*	கரிகால சோழனை *Karikāla cōḻaṉai* *Karikala Chola*
உலகில் முதல் தத்துவஞானி யார்? *Ulakil mutal tattuvañāṉi yār?* *Who was the first philosopher in the world?*	சாக்ரடீசு *Cākraṭīcu* *Socrates*	சாக்கிரட்டீசு *Cākkiraṭṭīcu* *Socrates*

4.2. Difference from Prior Work

This study contributes significantly to the realm of Question Answering (QA) systems, a pivotal area of research in Natural Language Processing (NLP). Notably, our exhaustive literature review revealed a scarcity of research focused specifically on the morphological analysis of answers within QA systems. However, we did encounter studies concentrating on morphology in isolation, which provided valuable insights that we leveraged to inform our research.

Upon identifying the research gap in applying morphological principles to QA systems, we endeavored to bridge this divide. By grounding our approach in the robust framework of Tamil grammar, as codified in the esteemed Nannūl, we successfully formulated a set of rigorous rules. These rules, rooted in the concrete grammatical structures of Tamil, enabled us to enhance the accuracy and efficacy of our QA system.

4.3. Areas of Concern Needing Attention

The integration of context-based analysis, translation, and transliteration techniques into the answer response processing pipeline is expected to effectively address challenges related to Named Entity Recognition, stop word removal, and management of multilingual content. By leveraging these advanced techniques, the accuracy and validity of the outcomes will be significantly enhanced. Specifically, resolving these issues is anticipated to improve the precision of entity identification, reduce linguistic and cultural biases, and increase the reliability of insights derived from the data. Ultimately, this will enable more informed decision-making and drive meaningful improvements in downstream applications.

5. CONCLUSION

This pioneering study unveiled the vast potential for enhancing the efficacy of evaluation metrics and models in Tamil question-answering systems, leveraging the seminal insights and recommendations presented herein. By synergistically integrating fundamental grammatical structures from the revered Tamil grammar book, Nannūl, our research yielded remarkable advancements, achieving an impressive 11.87% improvement in CHAII and 9.58% in SQuAD dataset. The significance of this study lies in its ability to demonstrate the importance of linguistically informed approaches in natural language processing, particularly for low-resource languages like Tamil. While acknowledging the potential limitations of our approach, particularly in handling complex linguistic phenomena such as idiomatic expressions and figurative language, we believe that our research provides a solid foundation for future investigations. The study's strengths are multifaceted, including its innovative approach, significant advancements in evaluation metrics, and potential for future research avenues, such as integrating Named Entity Recognition (NER), cross-lingual term handling, and stop word removal. However, we also recognize the need for future studies to address the challenges of adapting the methodology to other languages, exploring applicability to other natural language processing tasks, and further refining the approach to handle complex linguistic phenomena. In our opinion, this study demonstrates the power of interdisciplinary research, combining insights from linguistics and computer science to drive innovation in natural language processing. We believe that this work has the potential to inspire new research directions and applications, and we look forward to exploring these opportunities in future studies. By bridging the chasm between linguistic theory and computational models, we have unlocked the full potential of question-answering systems, poised to revolutionize the landscape of natural language processing.

Future studies can build upon our work by investigating the applicability of our methodology to other languages, particularly those with complex grammatical structures, and developing strategies to effectively handle idiomatic expressions, figurative language, and other complex linguistic phenomena. Additionally, exploring the integration of our approach with other natural language processing tasks, such as text summarization, sentiment analysis, and machine translation, can further advance the state-of-the-art in natural language processing. Conducting in-depth analyses to understand the impact of various linguistic features on the performance of question-answering systems and designing new metrics to accurately assess their performance, particularly in low-resource languages, are also promising research directions. By addressing these research directions, future studies can further advance the field and unlock new possibilities for natural language processing.

REFERENCES

[1] Zaib, M., Zhang, W., Sheng, Q.Z., Mahmood, A., & Zhang, Y. (2021). Conversational question answering: a survey. Knowledge and Information Systems, 64, 3151 - 3195.

[2] Allam, A.M., & Haggag, M.H. (2016). The Question Answering Systems: A Survey.

[3] Kolomiyets, O., & Moens, M. (2011). A survey on question-answering technology from an information retrieval perspective. Inf. Sci., 181, 5412-5434.

[4] Srba, I., & Bieliková, M. (2016). A Comprehensive Survey and Classification of Approaches for Community Question Answering. ACM Transactions on the Web (TWEB), 10, 1 - 63.

[5] Krishnan, A., Sriram, S.R., Ganesan, B., & Sridhar, S. (2023). An Extractive Question Answering System for the Tamil Language. Advances in Science and Technology, 124, 312 - 319.

[6] Murugathas, R., & Thayasivam, U. (2022). Domain-specific Question & Answer generation in Tamil. 2022 International Conference on Asian Language Processing (IALP), 323-328.

[7] Namasivayam, R.V., & Rajan, M. (2023). Answer Prediction for Questions from Tamil and Hindi Passages. Procedia Computer Science.

[8] Pandian, S.L., & Geetha, T.V. (2008). Tamil Question Classification Using Morpheme Features. GoTAL.

[9] A. K. M., A. K. C., V. Dhanalakshmi, R. U. Rekha, K. P. Soman and S. Rajendran, "Morphological Analyzer for Agglutinative Languages Using Machine Learning Approaches," 2009 International Conference on Advances in Recent Technologies in Communication and Computing, Kottayam, India, 2009, pp. 433-435, doi: 10.1109/ARTCom.2009.184.

[10] C.S., M., RK Rao, P., & Lalitha Devi, S. (2012). Tamil NER - Coping with Real Time Challenges.

[11] Ligozat, A., Grau, B., & Tribout, D. (2012). Morphological Resources for Precise Information Retrieval. International Conference on Text, Speech and Dialogue.

[12] Koehler, F., Schütze, H., & Atterer, M. (2008). A Question Answering System for German. Experiments with Morphological Linguistic Resources. International Conference on Language Resources and Evaluation.

[13] Benajiba, Y., Rosso, P., Abouenour, L., Trigui, O., Bouzoubaa, K., & Belguith, L.H. (2010). Question Answering. Handbook of Natural Language Processing.

[14] Bernhard, D., Cartoni, B., & Tribout, D. (2011). Évaluer la pertinence de la morphologie constructionnelle dans les systèmes de Question-Réponse (Evaluating the relevance of constructional morphology in question-answering systems). JEPTALNRECITAL.

[15] Kumar, M.A., & Dhanalakshmi, V. (2009). A Novel Approach to Morphological Analysis for Tamil Language.

[16] Thayaparan, M., Theivendiram, P., Megala, U., Nadarasamoorthy, N., Dias, G., Jayasena, S., & Ranathunga, S. (2016). Tamil Morphological Analyzer Using Support Vector Machines. International Conference on Applications of Natural Language to Data Bases.

[17] Rajasekar, M., & Geetha, A. (2021). Comparison of Machine Learning Methods for Tamil Morphological Analyzer. Intelligent Sustainable Systems.

[18] Akilan, R., & Naganathan, E.R. (2014). Morphological Analyzer for Classical Tamil Texts: A Rule- based approach.

[19] Hoque, M.M., Gonçalves, T., & Quaresma, P. (2013). Classifying Questions in Question Answering System Using Finite State Machines with a Simple Learning Approach. Pacific Asia Conference on Language, Information and Computation.

[20] Prins, R. (2005). Finite-state pre-processing for natural language analysis.

[21] Roche, E., & Shabes, Y. (1997). Finite-State Language Processing. Language, 75, 850.

[22] Hu, S., Zou, L., & Zhang, X. (2018). A State-transition Framework to Answer Complex Questions over Knowledge Base. Conference on Empirical Methods in Natural Language Processing.

[23] Puliyurkesigan. 2010-2020. "Pavaṇanti muṉivar iyaṟṟiya naṉṉūl, collatikāram, kāntikaiyurai, āṟumuka nāvalar", Chennai, Saratha Patippakam

[24] S. Lushanthan, A. R. Weerasinghe and D. L. Herath, "Morphological analyzer and generator for Tamil Language," 2014 14th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2014, pp. 190-196, doi: 10.1109/ICTER.2014.7083900. keywords: {Tamil Morphological Analyzer and Generator;Morphology;Finite State Transducer;Regular Expressions},

[25] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/1606.05250

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut and Mehryar Mohri, "OpenFst: A General and Efficient Weighted Finite-State Transducer Library", Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA 2007), volume 4783 of Lecture Notes in Computer Science, pages 11-23. Springer, 2007. http://www.openfst.org.

[26] Khanuja, Simran & Bansal, Diksha & Mehtani, Sarvesh & Khosla, Savya & Dey, Atreyee & Gopalan, Balaji & Margam, Dilip & Aggarwal, Pooja & Nagipogu, Rajiv Teja & Dave, Shachi & Gupta, Shruti & Gali, Subhash & Subramanian, Vish & Talukdar, Partha. (2021). MuRIL: Multilingual Representations for Indian Languages.

[27] Howard, A., Nathani, D., Thakkar, D., Elliott, J., Talukdar, P., & Culliton, P. (2021). chaii - Hindi and Tamil Question Answering. Retrieved from https://kaggle.com/competitions/chaii-hindi-and-tamil-question-answering