Dynamic Phrase Generation for Detection of Idioms of Gujarati Language using Diacritics and Suffix-based Rules

Jatin C. Modh[1]
Research Scholar
Gujarat Technological University
Ahmedabad, Gujarat, India

Jatinderkumar R. Saini[2]*
Professor and Director
Symbiosis Institute of Computer Studies and Research
Symbiosis International (Deemed University), Pune, India

***Abstract*—Gujarati is the language used for everyday communication in the state of Gujarat, India. The Gujarati language is also officially recognized by the constitution and the government of India. Gujarati script is based on the Devanagari script. An idiom is an expression, phrase, or word that has a different meaning from the literal meaning of the words in it. Idioms represent the cultural heritage of Gujarati language. Idioms are used in Gujarati language for effective communication and convey of an accurate message. No Machine Translation System does the accurate translation of Gujarati idioms to English or any other language. Different idiom phrases can be generated by adding diacritic(s) as well as suffix to the root or base form of the idiom. Many forms of single idiom make automatic idiom identification as well as machine translation more challenging. This paper focuses on the design and implementation of diacritics and suffix-based rules for dynamic phrase generation and detection of idioms of Gujarati language. This implementation helps in identifying Gujarati idiom present in any possible form in the Gujarati text. The obtained results with the execution of 7050 different Gujarati idiom phrases yield an accuracy of 99.73%. The results are encouraging enough to make the proposed implementation useful for Natural Language processing tasks related to Gujarati language idioms.**

***Keywords*—Diacritic; Gujarati; idiom; machine translation system (MTS); natural language processing (NLP); suffix; unicode transformation format (UTF)**

I. Introduction

Machine translation is the sub-field of Natural Language Processing (NLP) which is also a sub-field of Artificial intelligence (AI). Natural language processing is the study of any language by analyzing its structure and morphology. Natural language processing is challenging as different language has different grammatical structure. Vocabulary is important for the enrichment of the language. Idioms also contribute to the enrichment of the language. Idioms give impetus to the language. The idiom is an incomplete phrase as part of a sentence.

The people of Gujarat state are known as Gujarati. Idioms are the invaluable heritage of Gujarati language and for Gujarati people. Idiom in which a word or phrase becomes specific rather than its literal meaning. An idiom is in which a word or phrase becomes a specific meaning rather than its literal meaning. Gujarati idioms represent the customs, manners and beliefs of the people of Gujarat who speak Gujarati language. Gujarati idioms are the adornment of Gujarati language. Gujarati idioms are spoken in day-to-day communication and understood by every Gujarati.

When the speaker uses idioms, the listener is likely to mistakenly understand the literal meaning of the words if they do not already know the metaphor. Usually, idioms cannot be translated properly by any machine translation system. In most cases, the meaning changes when the idioms are translated into another language or it becomes misleading. In the Gujarati language, a particular idiom may have one or more forms or phrases. For the correct translation of Gujarati idioms into any other language, detection of all idiom forms or phrases is a very crucial task.

A. Gujarati Script

There are more than 46.1 million speakers of Gujarati language in the world. Gujarati is the 26th most spoken native language in the world [1]. Gujarati script is a script closely related to Devanagari script. It is a syllabic alphabet (abugida), in which every consonant carries the inherent vowel. Its principles are similar to Devanagari script principles [2]. It is distinguished from Devanagari script by not having a horizontal bar for its letter forms [3]. The Gujarati script is used to write the Gujarati language of Gujarat state in India. Gujarati language consist three different types of character: 34 independent consonants, 13 independent vowels and dependent vowel signs [3-5].

B. Gujarati Idioms

An idiom is a common phrase whose meaning is different from its literal meaning of the word. It is widely used and it has its popular meaning. For the correct translation of Gujarati idioms, identification of different forms of idioms from the input text is important. In Gujarati language, different and valid forms of idioms are possible by adding one or more specific diacritics marks to the base or root form of the Gujarati idiom. For example, હાથ આપ ‘hath aap’ is the base or root form of Gujarati idiom. It’s one meaning is “to help” in English language. From root form હાથ આપ, other valid idiom forms like હાથ આપવો ‘hath aapvo’, હાથ આપી ‘hath aapi’, હાથ આપીને ‘hath aapine’, હાથ આપ્યો ‘hath aapyo’, હાથ આપેલો ‘hath aapelo’ etc. can be generated. Identification of all

*Corresponding Author

Gujarati idiom phrases is concentrated here. Surrounding words are important for the idioms having more than one literal meaning in Gujarati language [6-9]. But the dynamic generation, as well as identification of all Gujarati idiom forms from the base form of the idiom, are focused here only.

C. Diacritics

Diacritic or accent is a mark which is added to a word that changes its function, sense or pronunciation. A diacritical mark is attached to a letter or word to show appropriate stress or sound [10]. Diacritic marks are more common in Gujarati language. Diacritics can be inserted below, before, after or above to a letter or word. In Gujarati language, every consonant carries the natural vowel. Diacritic vowel signs are added after, before, above, or below a consonant. Table I shows the diacritics for Gujarati language [11].

D. Gujarati Idioms with Diacritics

Gujarati language is rich in its morphological structure. Even single idiom has many possible forms. For example મેથીપાક આપ 'methipak aap' is the base form of bigram idiom. Its English translation is "to beat", but the same idiom can be used by many ways in the Gujarati sentences. After applying all diacritics to base idiom મેથીપાક આપ 'methipak aap', other possible forms are generated like મેથીપાક આપા, મેથીપાક આપિ, મેથીપાક આપી, મેથીપાક આપુ, મેથીપાક આપૂ, મેથીપાક આપે, મેથીપાક આપૈ, મેથીપાક આપો, મેથીપાક આપૌ, મેથીપાક આપં, મેથીપાક આપૃ, મેથીપાક આપેં, મેથીપાક આપોં, મેથીપાક આપ્, મેથીપાક આપઁ, મેથીપાક આપઃ, મેથીપાક આપૄ, મેથીપાક આપ૱ .But these many forms of idioms are not used in sentences. Commonly idiom forms used in sentences are મેથીપાક આપી, મેથીપાક આપે, મેથીપાક આપો .Other forms are also possible like મેથીપાક આપીને, મેથીપાક આપ્યો, મેથીપાક આપવા, મેથીપાક આપેલો. So rules are needed to generate different idiom phrases or forms that can be used in sentences.

E. Unicode Transformation Format (UTF)

The Unicode standard is based on ISCII-1988 (Indian Script Code for Information Interchange). The Unicode Standard encodes the Devanagari characters, as well as the same layout, is followed for Gujarati language [3]. Unicode or Universal Transformation Format is a variable width character encoding. It is capable of all valid code points in Unicode. It is a superset of American Standard Code of Information Interchange (ASCII). UTF goes beyond 8-bits and supports all languages in the world.

TABLE I. DIACRITICS FOR GUJARATI LANGUAGE

◌ા ā	◌િ i	◌ી ī	◌ુ u	◌ૂ ū
◌ે E	◌ૈ ai	◌ો o	◌ૌ au	◌ં ṁ
◌ૃ ṛ	◌ૅ Ē	◌ૉ Ō	◌ઃ	◌઼
◌ઃ ḥ	◌ૄ rī	◌ઁ		

The rest of the paper is organized as follows: Section II presents the literature review related to the study of Gujarati morphology, diacritic identification methods, stemmer, and Gujarati idioms; Section III covers the methodology in which diacritics and suffix based rules are generated for dynamic idiom generation; it also describes proposed algorithm steps to identify the different Gujarati idiom phrases within the input text. In Sections IV and V, experiments with results, analysis, and observations are discussed; finally, the conclusion part is discussed in Section VI.

II. RELATED LITERATURE REVIEW

Various projects have been carried out for the study of different languages and their machine translation, but the scope of this paper is Gujarati language and related to idioms only.

Rakholia et al. [11] implemented diacritic identification and extraction Technique for Gujarati language using an 8-bit Unicode Transformation Format. They designed an independent tokenization algorithm for diacritic extraction from Gujarati documents. They claimed 99.58% accuracy in diacritic extraction.

Sheth et al. [12] proposed a stemmer called Dhiya for morphological level analysis for Gujarati language. They used inflections of Gujarati text and created rule sets and tested stemmer performance using EMILLE corpus. They claimed 92.41% accuracy of stemmer.

Patel et al. [13] proposed a hybrid morphological analyzer paradigm model for Gujarati language. They applied a supervised approach and used partial stemmer for the generation of word forms based on language-dependent rules. They built a manual dictionary of root words, classified words and covered 5000 nouns.

Baxi et al. [14] developed a morphological analyzer for Gujarati language using a hybrid combination of knowledge-based, statistical-based and paradigm-based approach. They claimed 92.34% accuracy with a knowledge-based hybrid model and 82.84% accuracy with a statistical hybrid model.

Fashwan et al. [15] proposed rule-based method for detecting the case ending diacritics but it was for modern standard Arabic texts. They applied morphological analysis, part of speech, syntactic analysis and word relation as well as position. They experimented with both morphological and syntactic processing levels for handling diacritization problem.

Dan et al. [16] described the languages that use diacritical characters and difficulty in recovery of missing diacritics. They evaluated and described a system for automatically recovering the missing diacritics in documents in the Romanian language. They suggested recovery suggestions for possible changes for Romanian diacritics.

Rakholia et al. [17] implemented a rules-based technique to identify stop words from the Gujarati text. They presented 11 rules to identify a complete list of Gujarati stop words. They applied an automatic and dynamic approach to identify stop words from Gujarati documents and claimed 94.08% average accuracy.

Research works involving Natural Language Processing (NLP) of Gujarati language have been presented for MTS for Sanskrit-Gujarati pair [18], comparison of morphologically analyzed words [19], bilingual dictionary implementation [20], constituency mapper [21], classification [22] and information retrieval [23] to name a few.

Based on this literature review and the analysis based on Gujarati diacritics, no researchers have identified various idiom forms from the input text using the rule-based diacritics insertion technique. No researchers have applied diacritics and suffix based rules on idiom base form to generate all possible idioms. Some researchers have experimented on diacritization but using different languages other than the Gujarati language. Some of the researchers have applied various techniques for creating rule-based stemmer and diacritics identification methods.

The proposed model deals with the Gujarati idioms and their possible forms of idioms. Due to many phrases or forms of Gujarati idioms, the detection of Gujarati idioms within input text is a challenging task. The proposed model detects all Gujarati idioms present in the text by generating and searching all possible forms of particular idiom within the text. The proposed model applies dynamic phase generation for the detection of Gujarati idioms using diacritics and suffix-based rules. All available machine translation systems encounter problems in translating Gujarati idioms. Idiom detection helps the researchers' community in translating the Gujarati idioms into any language.

III. Methodology

In Gujarati language, distinct 3240 n-gram Gujarati idioms were collected. But in Gujarati language, one idiom can be used in many ways i.e. one specific idiom may have many forms or phrases. Rules are generated and applied on idiom base form to generate all possible forms of idioms. So for the generation of idiom forms, the base idiom form is stored in the database and all possible forms of idioms are generated dynamically by inserting diacritics and suffix to the base idiom form by applying defined rules. This implementation is used to identify any forms of Gujarati idioms within input Gujarati text.

A. Rules Generation

Rules are generated and applied for n-gram idioms where n>=2. For bigram or 2-gram idioms, rules are applied on the 2nd word only and various idiom forms can be generated. For trigram or 3-gram idioms, rules are applied on 3rd word only and many idiom forms can be generated. For example, હાથ આપ 'hath aap' is the bigram idiom root form, so diacritic rules are applied on 2nd word આપ 'aap' only; whereas અક્કલ મારી જવ 'akkal mari jav' is the trigram idiom root form, so diacritic rules are applied on 3rd word જવ 'jav' only. In general, for n-gram idiom where n>=2, then many idiom forms can be generated by applying rules on last word of idiom root form. For 1-gram idioms as well as some n-gram idiom(s), different forms of idioms are not applicable.

Following Rules are identified to generate possible idiom forms from the given base form of n-gram idiom.

1) Rule 0: Root or base form only

For instance, idiom અધ્ધર રાખ 'adhdhar rakh', the same form is used in sentences as an idiom. So no diacritics need to be added on root verb રાખ.

2) Rule 1: Root form + Diacritics

For instance, idiom અધ્ધર રાખ 'adhdhar rakh', root verb is રાખ 'rakh'. Based on Table II, after adding single diacritics, possible other forms of idiom અધ્ધર રાખ 'adhdhar rakh' are 18: અધ્ધર રાખા, અધ્ધર રાખિ, અધ્ધર રાખી, અધ્ધર રાખુ, અધ્ધર રાખૂ, અધ્ધર રાખે, અધ્ધર રાખૈ, અધ્ધર રાખો, અધ્ધર રાખૌ, અધ્ધર રાખં, અધ્ધર રાખૃ, અધ્ધર રાખૅ, અધ્ધર રાખૉ, અધ્ધર રાખ્, અધ્ધર રાખઁ, અધ્ધર રાખઃ, અધ્ધર રાખૄ, અધ્ધર રાખ .Out of these 18 generated forms, 05 common idiom forms used in Gujarati sentences are અધ્ધર રાખા, અધ્ધર રાખી, અધ્ધર રાખુ, અધ્ધર રાખે, અધ્ધર રાખો .Other 13 idioms forms are ignored as they are not used in general. By adding extra diacritics ◌̇; other 02 commonly used forms are generated as અધ્ધર રાખાં and અધ્ધર રાખું.

TABLE II. Example of Diacritics Insertion to Letter **ત** 'T' and Word **આગ લાગ** 'AAG LAG'

Sr. No	Diacritics	Example ત + Diacritics	Example આગ લાગ + **Diacritics**
1	◌ા	**તા**	**આગ લાગા**
2	◌િ	**તિ**	**આગ લાગિ**
3	◌ી	**તી**	**આગ લાગી**
4	◌ુ	**તુ**	**આગ લાગુ**
5	◌ૂ	**તૂ**	**આગ લાગૂ**
6	◌ે	**તે**	**આગ લાગે**
7	◌ૈ	**તૈ**	**આગ લાગૈ**
8	◌ો	**તો**	**આગ લાગો**
9	◌ૌ	**તૌ**	**આગ લાગૌ**
10	◌ં	**તં**	**આગ લાગં**
11	◌ૃ	**તૃ**	**આગ લાગૃ**
12	◌ૅ	**તૅ**	**આગ લાગૅ**
13	◌ૉ	**તૉ**	**આગ લાગૉ**
14	◌્	**ત્**	**આગ લાગ્**
15	◌ઁ	**તઁ**	**આગ લાગઁ**
16	◌ઃ	**તઃ**	**આગ લાગઃ**
17	◌ૄ	**તૄ**	**આગ લાગૄ**
18	◌઼	**त़**	**આગ લાग़**

3) Rule 2: using suffix વ and diacritics

Root form + વ + Diacritic(s)

Example: for idiom અધ્ધર રાખ 'adhdhar rakh', રાખ 'rakh' is the root verb form. Based on Table II, possible other forms of idiom અધ્ધર રાખ 'adhdhar rakh' are 18: અધ્ધર રાખવા, અધ્ધર રાખવિ, અધ્ધર રાખવી, અધ્ધર રાખવુ, અધ્ધર રાખવૂ, અધ્ધર રાખવે, અધ્ધર રાખવૈ, અધ્ધર રાખવો, અધ્ધર રાખવૌ, અધ્ધર રાખવં, અધ્ધર રાખવૃ, અધ્ધર રાખવૅ, અધ્ધર રાખવૉ, અધ્ધર રાખવ્, અધ્ધર રાખવઁ, અધ્ધર રાખવઃ, અધ્ધર રાખવૂ, અધ્ધર રાખવ .Out of these 18 generated forms, 04 common idiom forms used in Gujarati sentences are અધ્ધર રાખવા, અધ્ધર રાખવી, અધ્ધર રાખવુ, અધ્ધર રાખવો .Other idioms forms are ignored for considering rules generation.

By adding extra diacritics ં; other 02 commonly used forms are generated as અધ્ધર રાખવાં and અધ્ધર રાખવું.

Root form + વ + Diacritic ા + Diacritic ં

For idiom અધ્ધર રાખ,

અધ્ધર રાખ + વ + ા + ં = અધ્ધર રાખવાં

Root form + વ + Diacritic ુ + Diacritic ં

For idiom અધ્ધર રાખ,

અધ્ધર રાખ + વ + ુ + ં = અધ્ધર રાખવું

4) Rule 3: using suffix ય and diacritics

Root form + ્ + ય + Diacritic(s)

Example: for idiom અધ્ધર રાખ, રાખ is the root verb form. Common forms of idiom used in sentences are 5: અધ્ધર રાખ્યા, અધ્ધર રાખ્યાં, અધ્ધર રાખ્યુ, અધ્ધર રાખ્યું, અધ્ધર રાખ્યો

રાખ + ્ + ય + ા =રાખ્યા

રાખ + ્ + ય + ા + ં =રાખ્યાં

રાખ + ્ + ય + ુ =રાખ્યુ

રાખ + ્ + ય + ુ + ં =રાખ્યું

રાખ + ્ + ય + ો =રાખ્યો

5) Rule 4: using suffix ન and diacritics

Root form + Diacritic ી + ન + Diacritic ે

Example: for idiom સંસાર માંડ 'sansar mand', possible and common forms of idiom used in sentences are સંસાર માંડીને 'sansar mandine'

માંડ + ી + ન + ે = માંડીને

6) Rule 5: using suffix લ and diacritics

Root form + Diacritic ે + લ + Diacritic ા

Root form + Diacritic ે + લ + Diacritic ો

Example: for idiom અધ્ધર રાખ 'adhdhar rakh', Possible and common forms of idioms used in sentences are: અધ્ધર રાખેલા 'adhdhar rakhela', અધ્ધર રાખેલો 'adhdhar rakhelo'

રાખ + ે + લ + ા = રાખેલા

રાખ + ે + લ + ો = રાખેલો

7) Rule 6: using two suffixes and diacritics

Root form + વ + Diacritic ા + મ + Diacritic ા

Root form + વ + Diacritic ા + મ+ Diacritic ા + Diacritic ં

Example: for idiom અધ્ધર રાખ 'adhdhar rakh', Possible and common forms of idioms used in sentences are: અધ્ધર રાખવામા 'adhdhar rakhvama', અધ્ધર રાખવામાં 'adhdha rakhvaman'

રાખ + વ + ા + મ + ા = રાખવામા

રાખ + વ + ા + મ + ા + ં = રાખવામાં

8) Rule 7: using diacritic ૅ, suffix વ and diacritics

Root form + ે + વ + Diacritic(s)

Root form + ે + વ + ા

Root form + ે + વ + ી

Root form + ે + વ + ુ

Root form + ે + વ + ુ + ં

Root form + ે + વ + ો

If the last word of root form is રહ, then possible common forms of idioms used in sentences are રહેવા, રહેવી, રહેવુ, રહેવું, રહેવો

9) Rule 8: using diacritic ા, suffix and diacritics

Root form + ા + વ

Root form + ા + વ + ા

Root form + ા + વ + ા + ં

Root form + ા + વ + ી

Root form + ા + વ + ુ + ં

Root form + ા + વ + ો

Root form + ા + ય + ુ + ં

If the last word of root form is ભર 'bhar', then possible common forms of idioms used in sentences are ભરાવ, ભરાવા, ભરાવાં, ભરાવી, ભરાવું, ભરાવો, ભરાયું

10)Rule 9: using diacritic ૉ, suffix and diacritic(s)

For the last word of base idiom form is જ 'j'

Root form + ો + ઈ

Root form + ો + વ + ા

Root form + ો + વ + ા + ં

Root form + ો + વ + ી

Root form + ો + વ + ુ

Root form + ો + વ + ુ + ં

Root form + ો + વ + ો

If the last word of root form is જ, then possible common forms of idioms used in sentences are જોઈ, જોવા, જોવાં, જોવી, જોવુ, જોવું, જોવો

11)Rule 10: For the last word of base idiom form is જવ 'jav']Exception rule]

If the last word of root form is જવ, then possible common forms of idioms used in sentences are જવો, જવું, જવુ, જવી, જવાં, જવા, ગયો, ગયું, ગયુ, ગયાં, ગયા, ગઈ

12)Rule 11: For the last word of base idiom form is થવ 'thav' [Exception rule]

If the last word of root form is થવ, then possible common forms of idioms used in sentences are થવો, થવું, થવુ, થવી, થવાં, થવા, થયો, થયું, થયુ, થયાં, થયા, થઈ

13)Rule 12: For the last word of base idiom form is ખાવ 'khaav' [Exception rule]

If the last word of root form is ખાવ, then possible common forms of idioms used in sentences are ખાવો, ખાવું, ખાવુ, ખાવી, ખાવાં, ખાવા, ખાધી, ખાધાં, ખાધા

14)Rule 13: For the last word of base idiom form is લેવ 'lev' [Exception rule]

If the last word of root form is લેવ, then possible common forms of idioms used in sentences are લેવો, લેવું, લેવુ, લેવી, લેવાં, લેવા, લીધું, લીધુ, લીધાં, લીધા

15)Rule 14: For the last word of base idiom form is બેસ 'bes' [Exception rule]

If the last word of root form is બેસ, then possible common forms of idioms used in sentences are બેસી, બેસવો, બેસવું, બેસવુ, બેસવી, બેસવા, બેઠો, બેઠું, બેઠુ, બેઠી, બેઠાં, બેઠા

Table III shows all rules generated for inserting diacritics and suffix to root idiom form. Rule 0 is the original root form of idiom stored in the database. All rules from Rule 1 to Rule 9 are applied to all root forms of idioms. Rule 10 to Rule 14 are exception rules for idioms whose root form ends with જવ 'jav' થવ 'thav' ખાવ 'khaav' લેવ 'lev' બેસ 'bes'.

B. Proposed Model

To store the idiom database, MySQL was used for database software. PHP was used as a scripting language for the development platform. Visual Studio Code was used as an editor to write PHP coding. XAMP was used for the cross-platform local web server for the implementation of the model.

- Step 1: Data collection: A total of 3240 distinct Gujarati n-gram idioms were collected from different books and websites.
- Step 2: Pre-processing step: Except for 1-gram idioms, wherever applicable, the root form of each Gujarati idiom is generated; so diacritics and suffix can be added dynamically to the root form to generate various possible idiom forms.
- Step 3: Idiom database generated that contains idiom column in which root or base form of the idiom is stored once for each idiom. Idiom database contains root form idiom column with corresponding literal Gujarati meaning column for each idiom.
- Step 4: Accept input as text containing Gujarati idiom(s). Input text may contain any number of n-gram Gujarati idioms.
- Step 5: The algorithm searches all n-gram idioms from the input text by comparing input text with the idiom column of the idiom database.
- Step 6: Proposed algorithm generates all possible forms of idiom(s) for specific n-gram idiom. The algorithm uses the rules created and shown in Table III and generates all possible forms of idioms.
- Step 7: Generated various idiom form(s) are compared with the idiom form entered within the input text. If matching idiom form is found in the input text, the algorithm displays its possible literal meaning(s) in the Gujarati language; otherwise, it displays the text or idiom form as it is.

TABLE III. COMMON POSSIBLE IDIOM FORMS GENERATED USING RULE-0 TO RULE-14

Sr. No	Rule Description	Example for અધ્ધર રાખ 'adhdhar rakh'	Remarks
1.	Root Idiom	અધ્ધર રાખ	Rule 0
2.	Root Idiom + ◌ા	અધ્ધર રાખા	Rule 1
3.	Root Idiom + ◌ા + ◌ં	અધ્ધર રાખાં	
4.	Root Idiom + ◌ી	અધ્ધર રાખી	
5.	Root Idiom + ◌ુ	અધ્ધર રાખુ	
6.	Root Idiom + ◌ુ + ◌ં	અધ્ધર રાખું	
7.	Root Idiom + ◌ે	અધ્ધર રાખે	
8.	Root Idiom + ◌ો	અધ્ધર રાખો	
9.	Root Idiom + **વ** + ◌ા	અધ્ધર રાખવા	Rule 2
10.	Root Idiom + **વ** + ◌ા + ◌ં	અધ્ધર રાખવાં	
11.	Root Idiom + **વ** + ◌ી	અધ્ધર રાખવી	
12.	Root Idiom + **વ** + ◌ુ	અધ્ધર રાખવુ	
13.	Root Idiom + **વ** + ◌ુ + ◌ં	અધ્ધર રાખવું	
14.	Root Idiom + **વ** + ◌ો	અધ્ધર રાખવો	
15.	Root Idiom + ◌્ + **ય** + ◌ા	અધ્ધર રાખ્યા	Rule 3
16.	Root Idiom + ◌્ + **ય** + ◌ા + ◌ં	અધ્ધર રાખ્યાં	
17.	Root Idiom + ◌્ + **ય** + ◌ુ	અધ્ધર રાખ્યુ	
18.	Root Idiom + ◌્ + **ય** + ◌ુ + ◌ં	અધ્ધર રાખ્યું	
19.	Root Idiom + ◌્ + **ય** + ◌ો	અધ્ધર રાખ્યો	
20.	Root Idiom + ◌ી + **ન** + ◌ે	અધ્ધર રાખીને	Rule 4
21.	Root Idiom + ◌ે + **લ** + ◌ા	અધ્ધર રાખેલા	Rule 5
22.	Root Idiom + ◌ે + **લ** + ◌ો	અધ્ધર રાખેલો	
23.	Root Idiom+ **વ**+ ◌ા +**મ** + ◌ા	અધ્ધર રાખવામા	Rule 6
24.	Root Idiom+ **વ**+ ◌ા +**મ** + ◌ા + ◌ં	અધ્ધર રાખવામાં	
	Rule Description	Example for માનમાં રહ	**Remarks**
25.	Root Idiom + ◌ે + **વ** + ◌ા	માનમાં રહેવા	Rule 7
26.	Root Idiom + ◌ે + **વ** + ◌ી	માનમાં રહેવી	
27.	Root Idiom + ◌ે + **વ** + ◌ુ	માનમાં રહેવુ	
28.	Root Idiom + ◌ે + **વ** + ◌ુ + ◌ં	માનમાં રહેવું	
29.	Root Idiom + ◌ે + **વ** + ◌ો	માનમાં રહેવો	
	Rule Description	Example for ગળું ભર	**Remarks**
30.	Root Idiom + ◌ા + **વ**	ગળું ભરાવ	Rule 8
31.	Root Idiom + ◌ા + **વ** + ◌ા	ગળું ભરાવા	
32.	Root Idiom + ◌ા + **વ** + ◌ા + ◌ં	ગળું ભરાવાં	
33.	Root Idiom + ◌ા + **વ** + ◌ી	ગળું ભરાવી	
34.	Root Idiom + ◌ા + **વ** + ◌ુ + ◌ં	ગળું ભરાવું	

35.	Root Idiom + ◌ા + વ + ◌ો	ગળું ભરાવો	
36.	Root Idiom + ◌ા + ય + ◌ુ + ◌ં	ગળું ભરાયું	
	Rule Description	Example for પાણી જ	**Remarks**
37.	Root Idiom + ◌ો + ઈ	પાણી જોઈ	Rule 9
38.	Root Idiom + ◌ો + વ + ◌ા	પાણી જોવા	
39.	Root Idiom + ◌ો + વ + ◌ા + ◌ં	પાણી જોવાં	
40.	Root Idiom + ◌ો + વ + ◌ી	પાણી જોવી	
41.	Root Idiom + ◌ો + વ + ◌ુ	પાણી જોવુ	
42.	Root Idiom + ◌ો + વ + ◌ુ + ◌ં	પાણી જોવું	
43.	Root Idiom + ◌ો + વ + ◌ો	પાણી જોવો	
EXCEPTION RULES			
Sr. No	**Rule Description**	**Possible forms of idioms used in sentences**	**Remarks**
44.	If last word of Root Idiom is **જવ** 'jav'	જવો જવું જવુ જવી જવાં જવા ગયો ગયું ગયુ ગયાં ગયા ગઈ	Rule 10
45.	If last word of Root Idiom is **થવ** 'thav'	થવો થવું થવુ થવી થવાં થવા થયો થયું થયુ થયાં થયા થઈ	Rule 11
46.	If last word of Root Idiom is **ખાવ** 'khaav'	ખાવો ખાવું ખાવુ ખાવી ખાવાં ખાવા ખાધી ખાધાં ખાધા	Rule 12
47.	If last word of Root Idiom is **લેવ** 'lev'	લેવો લેવું લેવુ લેવી લેવાં લેવા લીધું લીધુ લીધાં લીધા	Rule 13
48.	If last word of Root Idiom is **બેસ** 'bes'	બેસી બેસવો બેસવું બેસવુ બેસવી બેસવા બેઠો બેઠું બેઠુ બેઠી બેઠાં બેઠા	Rule 14

IV. Results

There is no automated tool available to measure accuracy in Gujarati language. The help of 2 or 3 linguists was taken, particularly for manual verification of results. Obtained results are recorded manually side-by-side to calculate accuracy. Individual idiom with their possible and valid forms is analyzed and tested each form for accuracy. For example, different forms of root form તરસ ખાવ 'taras khav' and હાથ આપ 'hath aap' are tested. The literal meaning of Gujarati idiom તરસ ખાવ is દયા બતાવવી 'daya batavavi' in Gujarati and 'showing kindness' in English.

INPUT TEXT=તરસ ખાવી તરસ ખાવા

FINAL OUTPUT=દયા બતાવવી દયા બતાવવી

INPUT TEXT=હાથ આપ હાથ આપવો હાથ આપી હાથ આપીને હાથ આપ્યો હાથ આપેલો

FINAL OUTPUT=મદદ કરવી મદદ કરવી મદદ કરવી મદદ કરવી મદદ કરવી મદદ કરવી

For experiments, overall 7050 different valid idiom phrases or forms were entered as input text and tested for results. Only 19 idiom forms were not correctly identified, whereas 7031 idioms forms were correctly identified by proposed system. The accuracy obtained for the correct identification of the Gujarati idiom(s) from the Gujarati text was 99.73%. Idiom phrases or forms which were not correctly identified due to similarity in their root forms; for example જામી જવું 'jami javu' and જામી જવી 'jami javi' both are distinct idioms with distinct literal Gujarati meaning but their root forms are same જામી જવ 'jami jav'. Other error issues were due to inclusion of comma, hyphen, space, non-breaking space between words for n-gram idioms where n>=2.

V. Analysis and Discussion

Some observations, results and language ambiguities came out during the experimentation.

1) Applied algorithm generates possible forms from the root idiom stored in the database. For example, અન્ન પાણી ઝેર થઈ જવ 'anna pani zer thai jav' is the root form stored in the database and it generates various forms and successfully identifies all forms like અન્ન પાણી ઝેર થઈ જવા, અન્ન પાણી ઝેર થઈ જવાં, અન્ન પાણી ઝેર થઈ જવું, અન્ન પાણી ઝેર થઈ ગયા, અન્ન પાણી ઝેર થઈ ગયું etc entered as input.

2) However, all collected idioms i.e. idiom root forms are inserted in the database, if the idiom does not exist in the database then algorithm won't able to identify its various forms. Therefore the algorithm returns particular idiom form as it is in the text without its literal Gujarati meaning.

3) For generating various idiom forms, algorithm inserts diacritic ◌ં at the end of some idioms root forms and generates various idiom forms. For example, લેવુ and લેવું, લેવા and લેવાં, લીધુ and લીધું, લીધા and લીધાં etc are generated from the idiom root form *માથે લેવ* 'mathe lev'. This will help us in correcting any input spelling error of user and catches correctly all input idiom forms like *માથે* લેવુ and *માથે લેવું, માથે લેવા* and *માથે લેવાં, માથે લીધુ* and *માથે લીધું, માથે લીધા* and *માથે લીધાં.* and displays Gujarati literal meaning *જવાબદારી ઉઠાવવી* 'javabdari uthavavi'.

4) However, minimum and relevant rules are made; implemented algorithm may generate some general irrelevant idiom forms based on general examples. For example *અધ્ધર રાખ* 'adhdhar rakh' root form will generate *અધ્ધર રાખેવુ, અધ્ધર રાખેવું, અધ્ધર રાખેવો.* These extra forms are generated internally and it doesn't affect the output at all. Same rule will generate relevant forms for *માનમાં રહ* 'akkad rah' root form like *માનમાં રહેવુ, માનમાં રહેવું, માનમાં રહેવો.*

5) For 1-gram or unigram idioms, all idioms are stored as it is with its literal Gujarati meaning in the database. For example, *ઊંટવૈદ* 'untvaid', *ઝોડ* 'jhod', *ઢ* 'dh', *પ્રજ્ઞાચક્ષુ* 'pragnachakshu', *મખ્ખીચૂસ* 'makhkhichus', *લલ્લુ* 'lallu' etc are unigram idioms and are stored as it is in the database with its literal meaning in Gujarati language.

6) For n-gram idioms (n>=2), where root forms are irrelevant or only a single idiom form is possible, the same idiom phrase is entered in the idiom database. For example, *અટકળ પંચા દોઢસો* 'atkal pancha dodhso' is 3-gram idiom and its different forms are irrelevant in Gujarati language. So *અટકળ પંચા દોઢસો* 'atkal pancha dodhso' as it is stored in the idiom database instead of its root form. Another example, *અડિયલ ટટ્ટુ* 'adiyal tattu' is 2-gram idiom and its different forms are irrelevant in Gujarati language; so it is stored in the same form in the idiom database.

7) The algorithm additionally identifies nested idiom i.e. idiom within idiom from the input text. For example, *એક ઘાએ બે કટકા થવા* 'ek ghaae be katka thava' input produces intermediate output *તડ ને ફડ જવાબ થવો* 'tad ne fad javab thavo'; but *તડ ને ફડ* is also idiom and its literal meaning is *સ્પષ્ટ* 'spasht'; so for input text *એક ઘાએ બે કટકા થવા*, the final output is *સ્પષ્ટ જવાબ થવો* 'spasht javab thavo'.

An idiom database containing 3240 distinct idioms was created and tested 7050 different idiom phrases or forms. An implemented algorithm can find out all possible idiom phrases within Gujarati text by applying 15 rules of Table III on idiom root form whose root form or idiom is present in the idiom database. Spelling errors in Gujarati idioms can also be rectified by the proposed model.

VI. Conclusion

The proposed model that identifies all valid Gujarati idiom forms within Gujarati text and returns their literal Gujarati meaning was successfully implemented. Dynamic generation and identification of all Gujarati idiom phrases are focused here. By the exhaustive in-depth study of 3240 Gujarati idioms and their 7050 different idiom forms, 15 rules are generated. These rules are used to insert diacritic(s) and suffixes to the base or root form of Gujarati idiom. These dynamically generated different idiom forms are used to identify any idiom phrase inside the text. If the particular Gujarati idiom root form is not present in the database, then model returns the idiom as it is. So an entire assemblage of idioms is required for the success of the model. It is noteworthy that the proposed approach is not just diacritic based but also uses suffixes like વ, ય and ન.

Based on the results obtained from generating various possible idiom forms via rules implementation, it is advocated that the proposed rule-based system for generating various idiom forms is promising and worth implementation in the real world for the translation of Gujarati language idiom to any other language. Since Gujarati idioms are used in many forms in real life, all forms of idiom identification are challenging tasks for any machine translation system. The implemented rule-based system identifies most of the various forms of idioms. The proposed model opens the path for Gujarati idiom translation to any other language by finding all possible idioms forms within the input text. The task of context identification for multiple meanings idioms concerning the translation of Gujarati idioms is left for future scope.

References

[1] GujaratiLexicon, Gujaratilexicon.com, Available online: http://www.letslearngujarati.com/about-us (accessed May 23, 2021).

[2] Alejandro Gutman and Beatriz Avanzati, "Gujarati", The Language Gulper, Available online: http://www.languagesgulper.com/eng/Gujarati.html (accessed April 25, 2021).

[3] The Unicode® Standard Version 12.0 – Core Specification, Unicode, Inc., Available online: https://www.unicode.org/versions/Unicode12.0.0/ch12.pdf (accessed April 24, 2021).

[4] Wikipedia, "Gujarati language", https://en.m.wikipedia.org/wiki/Gujarati_language (accessed April 28, 2021).

[5] Xavier Nègre, "Multilingual Keyboard", Lexilogos 2002-2019, Available Online: https://www.lexilogos.com/keyboard/gujarati.htm (accessed April 28, 2021).

[6] Modh J. C. and Saini J. R., "Using IndoWordNet for Contextually Improved Machine Translation of Gujarati Idioms", International Journal of Advanced Computer Science and Applications(IJACSA), 12(1), 2021; Available online: http://dx.doi.org/10.14569/IJACSA.2021.0120128.

[7] Modh J. C. and Saini J. R., 2020, "Context Based MTS for Translating Gujarati Trigram and Bigram Idioms to English", 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-6, doi: 10.1109/INCET49848.2020.9154112; Available online: https://ieeexplore.ieee.org/document/9154112/.

[8] Modh J. C. and Saini J. R., 2018, "A Study of Machine Translation Approaches for Gujarati Language", International Journal of Advanced Research in Computer Science, Volume 9, No. 1, January-February 2018, pages 285-288; Available online: ijarcs.info/index.php/Ijarcs/article/download/5266/4497.

[9] Saini J. R. and Modh J. C., 2016, "GIdTra: A Dictionary-based MTS for Translating Gujarati Bigram Idioms to English", Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC) 22-24 December 2016, pages 192-196, Available online: https://ieeexplore.ieee.org/document/7913143.

[10] Nordquist and Richard, "Examples of Diacritical Marks." ThoughtCo., Available online: https://www.thoughtco.com/what-is-a-diacritic-mark-1690444 (accessed April 24, 2021).

[11] Rakholia R.M. and Saini J.R., 2015, "The Design and Implementation of Diacritic Extraction Technique for Gujarati Written Script using Unicode Transformation Format", proc. of IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT-2015), Coimbatore, India, vol. 2, pages 654-659, Available online: https://ieeexplore.ieee.org/document/7226037.

[12] Sheth J. and Patel B., "Dhiya: A stemmer for morphological level analysis of Gujarati language," 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014, pp. 151-154, doi: 10.1109/ICICICT.2014.6781269.

[13] Patel C. and Patel M., 2016, "Paradigm Model based Hybrid Morph Analyzer for Gujarati using Partial Stemmer", IJSRD - International Journal for Scientific Research & Development Vol. 3, Issue 11, 2016, Available online: http://www.ijsrd.com/articles/IJSRDV3I110454.pdf.

[14] Baxi J, Patel P. and Bhatt B.,2015, "Morphological Analyzer for Gujarati using Paradigm based approach with Knowledge based and Statistical Methods", Twelfth International Conference on Natural

Language Processing (ICON-2015), Available online: https://cdn.iiit.ac.in/cdn/ltrc.iiit.ac.in/icon2015/icon2015_proceedings/PDF/49_rp.pdf.

[15] Amany Fashwan and Sameh Alansary, 2017, "SHAKKIL: An Automatic Diacritization System for Modern Standard Arabic Texts", Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 84–93, Valencia, Spain, April 3, 2017, Available online: https://www.aclweb.org/anthology/W17-1311.pdf.

[16] Dan Tufiş and Alexandru Ceaușu, 2008, "DIAC+: A Professional Diacritics Recovering System", Institute for Artificial Intelligence, Romanian Academy, Available online: http://lrec-conf.org/proceedings/lrec2008/pdf/54_paper.pdf.

[17] Rakholia R.M. and Saini J.R., 2017, "A Rule-based Approach to Identify Stop Words for Gujarati Language", proc. of The 5th International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA-2016), Bhubaneshwar, India, vol. 515, pages 797-806, Available online: https://link.springer.com/chapter/10.1007/978-981-10-3153-3_79.

[18] Raulji J.K. and Saini J.R., "A Rule Based Architecture for Sanskrit to Gujarati Machine Translation System", proc. of International Conference on Emerging Trends in Engineering, Science and Technology (ICRISET-2018), Anand, India, in press with IEEE.

[19] Raulji J.K. and Saini J.R., "Sanskrit Stopword Analysis through Morphological Analyzer and its Gujarati Equivalent for MT System", proc. of International Conference on ICT for Sustainable Development (ICT4SD-2019), Panaji, Goa, India, in press with Springer.

[20] Raulji J.K. and Saini J.R., "Bilingual Dictionary for Sanskrit – Gujarati MT Implementation", proc. of International Conference on ICT for Sustainable Development (ICT4SD-2019), Panaji, Goa, India, in press with Springer.

[21] Raulji J.K. and Saini J.R., "Sanskrit-Gujarati Constituency Mapper for Machine Translation System", proc. of IEEE Bombay Section Signature Conference (IBSSC- 2019), Mumbai, India, in press with IEEE.

[22] Rakholia R.M. and Saini J.R., 2017, "Classification of Gujarati Documents using Naïve Bayes Classifier", Indian Journal of Science and Technology, vol. 10(5), pages 1-9, Available online: http://indjst.org/index.php/indjst/article/view/103233/78147.

[23] Rakholia R.M. and Saini J.R., 2017, "Information Retrieval for Gujarati Language using Cosine Similarity based Vector Space Model", proc. of The 5th International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA-2016), Bhubaneshwar, India, vol. 516, pages 1-9, Available online: https://link.springer.com/chapter/10.1007/978-981-10-3156-4_1.