A Novel Morphological Analysis based Approach for Dynamic Detection of Inflected Gujarati Idioms

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

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

Ketan Kotecha[3]
Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune, India

***Abstract*—The Gujarati language is primarily spoken by Gujarati people living in the state of Gujarat, India. It is the medium of communication in the state of Gujarat. In the Gujarati language, 'rudhiprayog' i.e. idiom is a very much popular form of expression. It represents the real flavour of the Gujarati language. The idiom is a group of words saying one thing literally but means something else when we explored it in context. Like Gujarati verbs, idioms can be written in many forms. Due to different morphological forms of the same Gujarati idiom, Gujarati idiom identification is a challenging task for any machine translation system (MTS). Accordingly many forms also make idiom translation more complicated. In the current paper, Gujarati idioms in their different inflected forms are collected, analyzed and classified based on ending words. After classifying idioms, their base or root forms are identified. Base idiom form and their possible idioms forms are morphologically analyzed and rules are generated based on the relationship between base form and possible inflected forms of idioms. These rules are used to generate possible idiom forms from the base idiom form. Gujarati idiom in any valid inflected form can be dynamically detected from the Gujarati input text using the proposed novel morphological analysis based approach. The results are encouraging enough to implement the proposed model of rules for natural language processing tasks as well as a machine translation system for Gujarati language idioms.**

Keywords*—Gujarati; idiom; machine translation system (MTS); morphological analysis; natural language processing (NLP); rudhiprayog

I. Introduction

Gujarati is the official language of Gujarat state of India and also recognized by the constitution of India. Gujarati is the Indo-Aryan branch of the Indo-European language family. It is spoken by more than 46 million people. Most of the people speaking Gujarati live in the Gujarat state. Also Gujarati communities spread in the UK, USA and all around the world. Gujarati language is used in newspapers, magazines, television, education, business, communication and in all type of media. Gujarati is written using the Devnagari script. In Gujarati language, two numbers singular and plural are used. Three genders masculine, feminine and neuters are used. Three cases nominative, oblique and locative are used. A Gujarati verb corresponds with a person, number and gender. These all are marked by suffixes attached to the verb root. These make Gujarati inflection fairly complex [1][2][3].

A. Gujarati Idioms

Gujarati idioms are used in many forms for day-to-day communication in Gujarati language. Idioms are used by Gujarati people for expressing feelings and thoughts. Machine translation of Gujarati idioms is important for the communication with the other non-Gujarati people. Idiom is the phrase whose collective meaning is not the same as literal meaning of its individual words. Gujarati idioms like any other language idioms cannot be translated literally.

B. Gujarati Idioms with Static Forms

Static idioms are the idioms where its different inflected forms are not possible. Static idioms can have only single form in Gujarati language. For example, અક્કલનો ઓથમીર 'akkalano othamira' (i.e. stupid) is the idiom in which other inflected form of idiom does not make any sense. Another Gujarati idiom example is એદીનો અખાડો 'edino akhado' (i.e. very lazy) where other inflected and valid idiom forms are not possible.

C. Gujarati Idioms with Inflected Forms

Inflected idioms are the idioms where its different morphological forms are possible. Infected idioms can be used by its different forms in the sentence. Inflected idioms are generally ended with Gujarati inflected verb form. Inflection can be applied on the last word of idiom that is usually base verb form. For example, ફાચર માર 'phacara mara' (i.e. to disrupt) is the base form of idiom. Its various inflected forms like ફાચર મારવી 'phacara maravi', ફાચર મારી 'phacara mari', ફાચર મારીને 'phacara marine', ફાચર મારવામાં 'phacara maravamam' are possible in the Gujarati sentence. Inflection is applied on the last word of base form i.e. on word માર here. Another example, ખટકો રાખ 'khatako rakha' (i.e. keep in mind) is the base form and its possible forms used in the sentences are ખટકો રાખ 'khatako rakha', ખટકો રાખવા 'khatako rakhava', ખટકો રાખવામા 'khatako rakhavama', ખટકો રાખવો 'khatako rakhavo', ખટકો રાખી 'khatako rakhi', ખટકો રાખીને 'khatako rakhine', ખટકો રાખેલા 'khatako rakhela', ખટકો રાખેલો 'khatako rakhelo', ખટકો રાખો 'khatako rakho', ખટકો રાખ્યો 'khatako rakhyo'; Inflected forms are applied on the last word રાખ 'rakha' here.

*Corresponding Author.

D. Stemming of Idioms

Stemming is the important process in natural language processing. The purpose of stemming is to standardize words to their common base form. Stemming removes the affixes of the word to get the root word or base word or stem word [4][5][6]. For example, ફાચર મારવી 'phacara maravi', ફાચર મારી 'phacara mari', ફાચર મારીને 'phacara marine' are various inflected idiom forms used in the sentences; stem/base form of these idioms form is ફાચર માર 'phacara mara' (i.e. to disrupt). Another example, ખટકો રાખ 'khatako rakha' (i.e. keep in mind) is the stem/base form of various inflected idiom forms like ખટકો રાખવા 'khatako rakhava', ખટકો રાખવામા 'khatako rakhavama', ખટકો રાખવો 'khatako rakhavo', ખટકો રાખી 'khatako rakhi', ખટકો રાખીને 'khatako rakhine' etc. So ફાચર માર 'phacara mara' (i.e. to disrupt) and ખટકો રાખ 'khatako rakha' (i.e. keep in mind) are the stem form of idioms. This stem/base form is stored once in the idiom database to recognize the idioms from the input text. Using this stem form, other inflected idioms forms can be generated.

The rest of the paper is organized as follows: Section II represents the literature review related to Gujarati idioms translation and its identification from the Gujarati text; Section III covers the methodology including idioms data collection and the steps of proposed model; Results and analysis are discussed in Section IV; finally conclusion, limitations and future work are discussed in Section V.

II. Related Literature Review

Various machine translation system projects have been carried out, especially for the English idiomatic text. Few projects have been taken out for non-English languages containing idioms.

Microsoft Translator and Google Translate are machine translation systems support Gujarati language translation also. Microsoft Translator [7][8] and Google Translate [9][10] use Microsoft cognitive services and Google Neural machine translation system respectively. Both machine translation systems are very much accepted but failed in the correct translation of Gujarati idiomatic text.

Authors [11] employed a GIdTra for translating only Gujarati bigram idioms into English language. They used dictionary based approach for identifying Gujarati idioms from the input text. They focused bigram idioms to test their approach. Other forms of bigram idiom cannot be identified.

Modh and Saini [12] discussed various machine translation approaches for the Gujarati language. The researchers have implemented a context-based Machine Translation System (MTS) for translating Gujarati bigram and trigram idioms into the English language [13].

Researchers [14] experimented n-gram model and used IndoWordNet for getting synonyms of surrounding words of particular idiom. They also exercised various context windows sizes in the case of ambiguity in finding meaning. They also worked on diacritics and suffix based rules [15].

Muzny and Zettlemoyer [16] applied a supervised approach for the automatic identification of English idioms from the corpus of Wiktionary multi-word definitions. They claimed 65% precision level of accuracy.

Verma and Vuppuluri [17] experimented with the combination of dictionary knowledge and web knowledge for English idiom identification. Authors claimed their approach as language as well as domain-independent. They also accepted about non-availability of the meanings of idiom phrases.

Hubers et al. [18] studied on whether the age and emigration length affects the knowledge of idiomatic expressions in the Dutch language or not. They concluded that emigration length negatively affects emigrant idiom knowledge.

Kessler and Friedrich [19] experimented on 9-to-10-year old children whether children can predict idiom-final words. The authors concluded that children rapidly activate multi-word units for idioms and decompose them only after a short delay.

Bakir and Umbu [20] studied American Pie movie script and analyzed the typology of idiomatic expression and contextual meanings of English idioms. Ramadhan et al. [21] prepared a report on the adoption of English movies in learning English idioms for English undergraduate students.

Based on this literature review and study, researchers dealing with idioms face problems in accurate understanding, identification and translation of idioms. Very less work is done especially for the analysis and translation of Gujarati idioms. No researchers have tried to recognize Gujarati idioms from the Gujarati inputted text. No researchers have analyzed various idiom forms of Gujarati language.

The paper focuses on the study of Gujarati idioms and its various forms used in the sentences. The scope of this paper is to generate rules from the analysis of Gujarati idioms collection with their possible forms and implementing these rules in the algorithm for finding all types of Gujarati idioms within inputted Gujarati text. This implementation helps in the simplification and translation of Gujarati idioms to any language in the world.

III. Methodology

Different Gujarati idioms are collected and categorized on the basis of static idioms and inflected idioms. Inflected idioms are again sub-categorized on the basis of different ending words. On the basis of analysis of collected Gujarati idioms, rules and base forms of idioms are generated. All base forms of idioms are stored in the idiom database. Finally, the idiom database and these rules are used in the dynamic generation of different forms of the same Gujarati idiom. This dynamic idiom generation helps in identifying all Gujarati idioms from the inputted text.

A. Data Collection

Overall 3410 distinct Gujarati idioms are collected from different Gujarati language sources [22][23]. From these idioms, 6047 valid different Gujarati idiom forms are

collected. For example, ખટકો રાખવો 'khatako rakhavo' idiom is collected; its distinct and base form is ખટકો રાખ 'khatako rakha' (i.e. keep in mind). From the base form, many inflection forms are possible by adding suffixes and diacritics but valid inflected forms used in Gujarati language are ખટકો રાખવા 'khatako rakhava', ખટકો રાખવો 'khatako rakhavo', ખટકો રાખી 'khatako rakhi', ખટકો રાખીને 'khatako rakhine', ખટકો રાખેલો 'khatako rakhelo', ખટકો રાખો 'khatako rakho', ખટકો રાખ્યો 'khatako rakhyo' etc. These inflected idioms are analyzed with their different possible forms for the derivation and finalization of the stemming rules for idioms. Base forms of all idioms are stored in the idiom database for further processing.

B. Idiom Classification

By analyzing the idioms collection, it is found that the idioms can be classified by three ways: (1) n-gram wise where n=1 to 9 (2) m-meaning wise idioms where m=1 to 7 (3) Static idioms and inflected idioms. In this paper, third classification is focused to generate the idiom identification rules. Static idioms are the idioms where only single idiom form is possible and therefore stemming is not appropriate. Inflected idioms are the idioms where various idiom forms from the base form can be generated. In the current paper, inflected idioms are analyzed. Various idioms with its inflected forms are analyzed to derive stem or base word form of particular idiom.

Table I shows the two types of idioms and its count. Static Gujarati idioms count is 215 and inflected Gujarati idioms count is 5832. Static idioms are the idioms where only single form of idiom is possible so stemming is not meaningful whereas inflected idioms are the idioms where stemming can be applied to derive base form or stem word. Inflected Gujarati idioms are usually ended with verb forms. This base verb form can generate other inflected verb forms of idioms by adding suffixes. For example, ફાચર મારવી 'phacara maravi', ફાચર મારી 'phacara mari', ફાચર મારીને 'phacara marine' are the inflected idioms; it is ended with the different verb forms of માર 'mara'; so base form of these idioms is ફાચર માર 'phacara mara'.

Inflected idioms can further be classified on the base of end words. Inflected idioms can be of four categories on the base of end character(s) or word. (1) Idioms end with વું 'vum' (2) Idioms end with વા 'va' or વાં 'vam' (3) Idioms end with વી 'vi' (4) Idioms end with વો 'vo'. Table II shows these four categories of idioms and their counts.

TABLE I. TYPES OF GUJARATI IDIOMS WITH REFERENCE TO STEMMING

Sr. No.	Types of Idioms	Count	Remarks
1	Static idioms	215	Not possible to derive stem/base word. Only single form of idiom is possible. No similar structure in idioms. Example એદીનો અખાડો 'edino akhado' i.e. very lazy, બારમો ચંદ્રમા 'baramo candrama' i.e. animosity, આડી વાટની ધૂળ 'adi vatani dhula' i.e. fruitless work, ઓઢવા આભ ને પાથરવા ધરતી 'odhava abha ne patharava dharati' i.e. very miserable situation, ગાંઠનું ગોપીચંદન 'ganthanum gopicandana' i.e. at own cost
2	Inflected idioms	5832	Stem word or base form can be derived. Other idiom forms can be generated from the base form. Idiom is usually ending with the base verb form. Example ફાચર મારવી 'phacara maravi'→ ફાચર માર 'phacara mara' ખટકો રાખવો 'khatako rakhavo'→ ખટકો રાખ 'khatako rakha' Here ફાચર માર phacara mara' is the base form of idiom ફાચર મારવી 'phacara maravi' and માર 'mara' is verb form.
	Total	6047	

TABLE II. INFLECTED IDIOMS WHERE STEMMING CAN BE APPLIED

Sr. No.	Inflected Idioms end with the word	Count	Example
1	વું 'vum'	2534	ગોથું ખાવું 'gothum khavum' i.e. make a mistake
2	વા OR વાં 'va' 'vam'	741	આકડા વાવવા 'akada vavava' i.e. planting the roots of animosity ઝાવાં નાખવાં 'jhavam nakhavam' i.e. to boggle
3	વી 'vi'	1370	આંખ ઠરવી 'ankha tharavi' i.e. to be satisfied
4	વો 'vo'	1187	ખોળો ખાલી હોવો 'kholo khali hovo' i.e. to be childless
	Total	5832	

Category 1 Idioms end with word વું : All idioms end with word વું are studied and can be further classified as shown in Table III; Idioms end with word એવું are 18, idioms end with word કવું are 68, idioms end with word કાવું are 2, idioms end with word ખવું are 133, idioms end with word ખાવું are 36, idioms end with word ગવું are 52 and so on. Corresponding base form of these idioms are derived as એવ, મૂક, સુક, રાખ, ખાવ, લાગ and so on. Table III shows a snapshot of partial data for the sub-categories of idioms end with word વું.

TABLE III. SNAPSHOT OF PARTIAL DATA OF IDIOMS END WITH THE WORD વું 'VUM'

Sr No.	Idioms end with the word	Count	Base or stem word derivation
1	એવું 'evum'	18	એવું 'evum'→ એવ 'eva'
2	કવું 'kavum'	68	મૂકવું 'mukavum'→ મૂક 'muka'
3	કાવું 'kaavum'	2	સુકાવું 'sukavum'→ સુક 'suka'
4	ખવું 'khavum'	133	રાખવું 'rakhavum'→ રાખ 'rakha'
5	ખાવું 'khaavum'	36	ખાવું 'khavum'→ ખાવ 'khava'
6	ગવું 'gavum'	52	લાગવું 'lagavum'→ લાગ 'laga'
7	ચવું 'chavum'	10	રચવું 'racavum'→ રચ 'raca'
8	ચાવું 'chaavum'	1	ગૂંચાવું 'guncavum'→ ગૂંચ 'gunca'
9	છવું 'chhavum'	6	પૂછવું 'puchavum' → પૂછ 'pucha'
10	જવું 'javum'	289	આંજવું 'anjavum '→ આંજવ 'anjava'

Category 2 Idioms end with word વા or વાં: All idioms end with word વા or વાં are studied and can be further classified as shown in Table IV; Idioms end with word ઝવા are 1, idioms end with word ટવા are 6, idioms end with word ટવાં are 2, idioms end with word ઠવા are 4 and so on. Corresponding base form of these idioms are derived as સૂઝ, પલટ, વાટવ, ઊઠ and so on. Table IV shows a snapshot of partial data for the sub-categories of idioms end with word વા or વાં.

TABLE IV. SNAPSHOT OF PARTIAL DATA OF IDIOMS END WITH THE WORD વા 'VA' OR વાં 'VAM'

Sr No.	Idioms end with the word	Count	Base or stem word derivation
1	ઝવા 'jhava'	1	સૂઝવા 'sujhava'→ સૂઝ 'sujha'
2	ટવા 'tava'	6	પલટવા 'palatava' → પલટ 'palata'
3	ટવાં 'tavam'	2	વાટવાં 'vatavam'→ વાટવ 'vatava'
4	ઠવા 'thava'	4	ઊઠવા 'uthava'→ ઊઠ 'utha'
5	ઠવાં 'thavam'	4	ઊઠવાં 'uthavam'→ ઊઠ 'utha'
6	ડવા 'dava'	57	ઊડવા 'udava'→ ઊડ 'uda'
7	ડવાં 'davam'	32	પડવાં 'padavam'→ પડ 'pada'
8	ઢવા 'dhava'	12	કાઢવા 'kadhava'→ કાઢ 'kadha'
9	ણવા 'nava'	11	ગણવા 'ganava'→ ગણ 'gana'
10	થવા 'thavam'	16	થવા 'thava'→ થવ 'thav'

Category 3 Idioms end with word વી: All idioms end with word વી are studied and can be further classified as shown in Table V; Idioms end with word તવી are 3, idioms end with word થવી are 33, idioms end with word દવી are 5, idioms end with word દેવી are 21, idioms end with word ધવી are 27 and so on. Corresponding base form of these idioms are derived as જીત, થવ, ખોદ, દેવ, બાંધ and so on. Table V shows a snapshot of partial data for the sub-categories of idioms end with word વી.

TABLE V. SNAPSHOT OF PARTIAL DATA OF IDIOMS END WITH THE WORD વી 'VI'

Sr No.	Idioms end with the word	Count	Base or stem word derivation
1	તવી 'tavi '	3	જીતવી 'jitavi'→જીત 'jita'
2	થવી 'thavi'	33	થવી 'thavi'→ થવ 'thava'
3	દવી 'davi'	5	ખોદવી 'khodavi'→ખોદ 'khod'
4	દેવી 'devi'	21	દેવી 'devi'→ દેવ 'deva'
5	ધવી 'dhavi'	27	બાંધવી 'bandhavi'→બાંધ 'bandha'
6	પવી 'pavi'	32	ચાંપવી 'campavi' →ચાંપ 'campa'
7	બવી 'bavi'	4	દાબવી 'dabavi' →દાબ 'daba'
8	મવી 'mavi'	12	રમવી 'ramavi'→રમ 'rama'
9	રવી 'ravi'	269	મારવી 'maravi'→માર 'mara'
10	લવી 'lavi'	58	ખોલવી 'kolavi'→ખોલ 'khola'

Category 4 Idioms end with word વો: All idioms end with word વો are studied and can be further classified as shown in Table VI; Idioms end with word ડવો are 138, idioms end with word પવો are 51, idioms end with word મવો are 19, idioms end with word રવો are 183, idioms end with word લવો are 31 and so on. Corresponding base form of these idioms are derived as પાડ, કાપ, જામ, ભર, ખેલ and so on. Table VI shows a snapshot of partial data for the sub-categories of idioms end with word વો.

TABLE VI. SNAPSHOT OF PARTIAL DATA OF IDIOMS END WITH THE WORD વો 'VO'

Sr No.	Idioms end with the word	Count	Base or stem word derivation
1	ડવો 'davo'	138	પાડવો 'padavo'→પાડ 'pada'
2	પવો 'pavo'	51	કાપવો 'kapavo'→કાપ 'kapa'
3	મવો 'mavo'	19	જામવો 'jamavo'→જામ 'jama'
4	રવો 'ravo'	183	ભરવો 'bharavo'→ભર 'bhara'
5	લવો 'lavo'	31	ખેલવો 'kheavo'→ખેલ 'khela'
6	લેવો 'levo'	27	લેવો 'levo'→ લેવ 'leva'
7	વવો 'vavo'	139	આવવો 'avavo'→આવ 'ava'
8	સવો 'savo'	11	ખસવો 'khasavo'→ ખસ 'khasa'
9	હોવો 'hovo'	39	હોવો 'hovo'→ હોવ 'hova'
10	ળવો 'lavo'	53	ઢોળવો 'dholavo'→ઢોળ 'dhola'

C. Rules Derivation [for Dynamic Inflected Idioms Generation from the base form]

By studying and analyzing all four categories of inflected idioms specified in Table III to Table VI, base forms of all idioms are collected. Further, by detailed morphological analysis and using reverse rules generation process, rules are defined to generate various idiom forms from the given base form of particular idiom as shown in Table VII.

Table VII shows the collection of rules for generating dynamic inflected idiom forms from the idiom base form by adding diacritics as well as different suffix characters. For example: ખટકો રાખ 'khatako rakha' is the base idiom form as per Rule 1. As per rule 2, second inflected form ખટકો રાખવા 'khatako rakhava' can be generated by adding suffix 'વ' and diacritics 'ા' to base idiom ખટકો રાખ 'khatako rakha'. As per rule 3, third inflected form ખટકો રાખવાં 'khatako rakhavam' can be generated by adding suffix 'વ' and two diacritics 'ા', 'ં' to base idiom ખટકો રાખ 'khatako rakha'. similar way using remaining rules overall 43 different inflected idioms forms can be generated. All other inflected idioms forms can be generated by adding different character વ 'va' ય 'ya' ન 'na' લ 'la' ઈ 'i' as well as by adding different diacritics as shown in Table VII.

TABLE VII. RULES TO GENERATE POSSIBLE FORMS OF IDIOMS FROM THE BASE IDIOM

Rule No.	Rules Definition	Base form	Inflected idiom forms
1	Base idiom	ખટકો રાખ	ખટકો રાખ
2	Base idiom+'વ'+'ા'	ખટકો રાખ	ખટકો રાખવા
3	Base idiom+'વ'+'ા'+'ં'	ખટકો રાખ	ખટકો રાખવાં
4	Base idiom+'વ'+'ા'+'મ'+'ા'	ખટકો રાખ	ખટકો રાખવામા
5	Base idiom+'વ'+'ા'+'મ'+'ા'+'ં'	ખટકો રાખ	ખટકો રાખવામાં
6	Base idiom+'વ'+'ી'	ખટકો રાખ	ખટકો રાખવી
7	Base idiom+'વ'+'ુ'	ખટકો રાખ	ખટકો રાખવુ
8	Base idiom+'વ'+'ુ'+'ં'	ખટકો રાખ	ખટકો રાખવું
9	Base idiom+'વ'+'ો'	ખટકો રાખ	ખટકો રાખવો
10	Base idiom+'ા'	ખટકો રાખ	ખટકો રાખા
11	Base idiom+'ા'+'ં'	ખટકો રાખ	ખટકો રાખાં
12	Base idiom+'ા'+'ય'+'ુ'+'ં'	ખટકો રાખ	ખટકો રાખાયું
13	Base idiom+'ા'+'વ'	ખટકો રાખ	ખટકો રાખાવ
14	Base idiom+'ા'+'વ'+'ા'	ખટકો રાખ	ખટકો રાખાવા
15	Base idiom+'ા'+'વ'+'ા'+'ં'	ખટકો રાખ	ખટકો રાખાવાં
16	Base idiom+'ા'+'વ'+'ી'	ખટકો રાખ	ખટકો રાખાવી
17	Base idiom+'ા'+'વ'+'ુ'+'ં'	ખટકો રાખ	ખટકો રાખાવું
18	Base idiom+'ા'+'વ'+'ો'	ખટકો રાખ	ખટકો રાખાવો
19	Base idiom+'ી'	ખટકો રાખ	ખટકો રાખી
20	Base idiom+'ી'+'ન'+'ે'	ખટકો રાખ	ખટકો રાખીને
21	Base idiom+'ુ'	ખટકો રાખ	ખટકો રાખુ
22	Base idiom+'ુ'+'ં'	ખટકો રાખ	ખટકો રાખું
23	Base idiom+'ે'	ખટકો રાખ	ખટકો રાખે
24	Base idiom+'ે'+'લ'+'ા'	ખટકો રાખ	ખટકો રાખેલા
25	Base idiom+'ે'+'લ'+'ો'	ખટકો રાખ	ખટકો રાખેલો
26	Base idiom+'ે'+'વ'+'ા'	ખટકો રાખ	ખટકો રાખેવા
27	Base idiom+'ે'+'વ'+'ી'	ખટકો રાખ	ખટકો રાખેવી
28	Base idiom+'ે'+'વ'+'ુ'	ખટકો રાખ	ખટકો રાખેવુ
29	Base idiom+'ે'+'વ'+'ુ'+'ં'	ખટકો રાખ	ખટકો રાખેવું
30	Base idiom+'ે'+'વ'+'ો'	ખટકો રાખ	ખટકો રાખેવો
31	Base idiom+'ો'	ખટકો રાખ	ખટકો રાખો
32	Base idiom+'ો'+'ઈ'	ખટકો રાખ	ખટકો રાખોઇ
33	Base idiom+'ો'+'વ'+'ા'	ખટકો રાખ	ખટકો રાખોવા
34	Base idiom+'ો'+'વ'+'ા'+'ં'	ખટકો રાખ	ખટકો રાખોવાં
35	Base idiom+'ો'+'વ'+'ી'	ખટકો રાખ	ખટકો રાખોવી
36	Base idiom+'ો'+'વ'+'ુ'	ખટકો રાખ	ખટકો રાખોવુ
37	Base idiom+'ો'+'વ'+'ુ'+'ં'	ખટકો રાખ	ખટકો રાખોવું
38	Base idiom+'ો'+'વ'+'ો'+' +	ખટકો રાખ	ખટકો રાખોવો
39	Base idiom+'્'+'ય'+'ા'	ખટકો રાખ	ખટકો રાખ્યા
40	Base idiom+'્'+'ય'+'ા'+'ં'	ખટકો રાખ	ખટકો રાખ્યાં
41	Base idiom+'્'+'ય'+'ુ'	ખટકો રાખ	ખટકો રાખ્યુ
42	Base idiom+'્'+'ય'+'ુ'+'ં'	ખટકો રાખ	ખટકો રાખ્યું
43	Base idiom+'્'+'ય'+'ો'	ખટકો રાખ	ખટકો રાખ્યો

Rules definitions specified in Table VII are applied on the base idiom form to generate all possible idiom forms dynamically. These rules help in searching any inflected idiom form available in the input text.

D. Idiom Database Creation

Database of idioms is mainly required to store the distinct base form of idiom and its simplified Gujarati meaning. Idiom database is created with idiom column and other related columns like Gujarati meaning, English meaning etc. Static idioms are having single form so they are stored in idiom column as it is. For inflected idioms, their base form is stored once in the idiom column. Gujarati meaning column stores the meaning of particular Gujarati idiom in simple Gujarati words.

E. Proposed Model

Fig. 1 shows the algorithm steps for the proposed model.

Step-1: Input the Gujarati text OR Input the file with encoding format UTF-8

Step-2: Apply preprocessing steps on it

2.1:Remove extra spaces from the both end of text

2.2:Remove all whitespace spaces in between the text

2.3:Remove special characters (\ , " < > . ? ; : !)

Step-3: Tokenize the words of input text

Step-4: Search idioms from the input text by implementing rules generated in Table VII and comparing with idiom column of idiom database.

Step-5: Display result (all idiom forms detected and replaced with Gujarati meaning)

Fig. 1. Algorithm for the Proposed Model.

Input is the Gujarati idiomatic text that may contain any number of Gujarati idioms. Output will be the Gujarati text without any idiom. In the output text, all idioms will be replaced with the value of Gujarati meaning column. Other Gujarati text will remain as it is. Output is nothing but Gujarati text without any idioms. For example,

INPUT TEXT= કોઈ પણ કામમાં ફાયર મારવી એ તેનો સ્વભાવ છે. 'ko'i pana kamamam phacara maravi e teno svabhava che'.

FINAL OUTPUT= કોઈ પણ કામમાં વચમાં અડચણ નાખવી એ તેનો સ્વભાવ છે. 'ko'i pana kamamam vacamam adacana nakhavi e teno svabhava che'.

IV. Result and Analysis

Overall 7400 different idiom forms were entered as input text. Input text may contain one or more idiom phrases. Output results were obtained. The correctness of output results was verified by the two linguists with Gujarati mother tongue and doctorate degree in Gujarati language. The proposed model was able to detect all inflected idiom forms present within Gujarati text successfully. The input text is nothing but the Gujarati idiomatic text using any inflected idiom form and the output shows the replacement of Gujarati idiom with simple Gujarati meaning of the particular idiom. For example;

Example 1:

INPUT TEXT= ખટકો રાખવા ખટકો રાખી

'khatako rakhava khatako rakhi'

FINAL OUTPUT= કોઈ વાત બરાબર ધ્યાનમાં રાખવી કોઈ વાત બરાબર ધ્યાનમાં રાખવી.

'ko'i vata barabara dhyanamam rakhavi ko'i vata barabara dhyanamam rakhavi'.

Example 2:

INPUT TEXT= વિદ્યાર્થીઓએ અભ્યાસ માટે ખટકો રાખવો જરૂરી બને છે. 'vidyarthi'o'e abhyasa mate khatako rakhavo jaruri bane che'.

FINAL OUTPUT= વિદ્યાર્થીઓએ અભ્યાસ માટે કોઈ વાત બરાબર ધ્યાનમાં રાખવી જરૂરી બને છે. 'vidyarthi'o'e abhyasa mate ko'i vata barabara dhyanamam rakhavi jaruri bane che'.

For understanding, ખટકો રાખ 'khatako rakha' (i.e. keep in mind) is the base form of idiom. Its many inflection forms are possible by adding suffix વ ય ન લ ઈ and/or by adding different diacritics. But only valid different forms are given for testing like ખટકો રાખવા, ખટકો રાખવામા, ખટકો રાખવો, ખટકો રાખી, ખટકો રાખીને, ખટકો રાખે, ખટકો રાખેલા, ખટકો રાખેલો, ખટકો રાખો, ખટકો રાખ્યા, ખટકો રાખ્યો; Proposed algorithm detects all the inflected forms of base form idiom ખટકો રાખ and so algorithm displays the output as Gujarati meaning of idiom as કોઈ વાત બરાબર ધ્યાનમાં રાખવી. Other Gujarati text will remain as it is in the output. Table VIII displays Output text for the given Input text. Output text is the same Gujarati meaning for any inflected idiom form of the same base idiom.

TABLE VIII. Different Inflected Forms of Idiom as Input and Corresponding Output

Sr No.	Input text	Output text (Gujarati meaning)
1	ખટકો રાખ 'khatako rakha'	કોઈ વાત બરાબર ધ્યાનમાં રાખવી 'ko'i vata barabara dhyanamam rakhavi'
2	ખટકો રાખવા 'khatako rakhava'	
3	ખટકો રાખવામા 'khatako rakhavama'	
4	ખટકો રાખવો 'khatako rakhavo'	
5	ખટકો રાખી 'khatako rakhi'	
6	ખટકો રાખીને 'khatako rakhine'	
7	ખટકો રાખે 'khatako rakhe'	
8	ખટકો રાખેલા 'khatako rakhela'	
9	ખટકો રાખેલો 'khatako rakhelo'	
10	ખટકો રાખો 'khatako rakho'	
11	ખટકો રાખ્યા 'khatako rakhya'	
12	ખટકો રાખ્યો 'khatako rakhyo'	
13	ફાયર માર 'phacara mara'	વચમાં અડચણ નાખવી 'vacamam adacana nakhavi'
14	ફાયર મારવી 'phacara maravi'	
15	ફાયર મારી 'phacara mari'	
16	ફાયર મારીને 'phacara marine'	
17	ફાયર મારવામાં 'phacara maravamam'	
18	ભાંગરો વાટ 'bhangaro vata'	છૂપી વાત ખુલ્લી કરવી 'chhupi vata khulli karavi'
19	ભાંગરો વાટી 'bhangaro vati'	
20	ભાંગરો વાટવો 'bhangaro vatavo'	
21	ભાંગરો વાટ્યો 'bhangaro vatyo'	
22	ભાંગરો વાટીને 'bhangaro vatine'	

Features of proposed algorithm are as follows:

1) Applied algorithm is domain independent. Proposed implementation detects all Gujarati inflected idioms used anywhere in the input text and replaces all Gujarati idioms with simple Gujarati meaning.

2) Proposed model stores base form of inflected idiom in the database as single record. For example base form ભાંગરો વાટ 'bhangaro vata' (i.e. disclose a secret) is stored in the idiom database once, but it is used to generate all possible forms of the same idiom like ભાંગરો વાટી 'bhangaro vati', ભાંગરો વાટવો 'bhangaro vatavo', ભાંગરો વાટ્યો 'bhangaro vatyo', ભાંગરો વાટીને 'bhangaro vatine' etc.

3) Dictionary based approach is applied for searching static idioms from the input text because static idioms are having the single possible idiom form. No rules are applicable on static idioms as they are found in irregular forms like અક્કલની ખાણ 'akkalani khana' i.e. very intelligent person, આંખનો પાટો 'ankhano pato' i.e. disgusting, કાચું સોનું 'kacum sonum' i.e. very fertile land, ખાસડાને તોલે 'khasadane tole' i.e

inferior, બારે દહાડા ને બત્રીસે ઘડી 'bare dahada ne batrise ghadi' i.e. persistent, માથે રાત જેવું ધાબું 'mathe rata jevum dhabum' i.e. utter darkness, વરઘોડાની વાડી 'varaghodani vadi' i.e. transient, વાડ તેવો વેલો 'vada tevo velo' i.e. people follows king etc.

4) Inflected idioms are generally ended with the words વું 'vum', વા 'va', વાં 'vam', વી 'vi', વો 'vo'. These idioms are collected and analyzed in Table III to Table VI. Considering limited number of idioms, base idiom forms of these idioms are collected manually. Using reverse process, algorithm is developed to generate all idioms forms from the base form. Only base forms of idioms are stored once in the idiom database.

5) Proposed algorithm generates all possible forms of idioms by applying all Table VII rules to base idiom form for detection of any idiom in the input text. So it sometimes rectifies minor spelling mistakes in the Gujarati idiom form automatically. For example, ભાંગરો વાટ્ય 'bhangaro vatya' and ભાંગરો વાટવ 'bhangaro vatava' both are erroneous spelling forms of base idiom ભાંગરો વાટ 'bhangaro vata' i.e. disclose a secret; but the proposed algorithm considers and corrects both as ભાંગરો વાટ 'bhangaro vata' i.e. disclose a secret.

6) The proposed model is the first approach in Gujarati language that able to find out any valid and possible forms of Gujarati idioms present in the Gujarati text and provides Gujarati simplification of the particular idiom.

V. Conclusion, Limitations and Future Work

The proposed rule-based model was successfully implemented and it successfully detected all the static and inflected Gujarati idiom forms from the Gujarati text. The proposed algorithm successfully detected all the idioms by implementing a dictionary-based approach as well as dynamic idiom form generation rule-based approach. The proposed algorithm generates all possible idiom forms dynamically to determine whether any inflected form of a particular base form idiom is present in the input text or not.

The proposed system can detect any form of the idiom from the Gujarati text but the thing is that the particular idiom base form must be present in the idiom database. The proposed system could not identify idiom that is not available in the idiom database. Future work is to collect all Gujarati idioms from possible sources to rectify this shortcoming. Also algorithm applies all the generated rules to all base form of idioms for generating possible idiom forms.

Based on the results obtained, it is advocated that the proposed system is worth implementation in the real world for machine translation of the Gujarati language. All pioneering machine translation systems for the Gujarati language including Google Translate and Microsoft Translator face the problem of idiom translation. The proposed system successfully identifies the Gujarati idioms available in the input text. The idiom identification method of the proposed model makes it easy for any machine translation system to deal with the Gujarati idiom. Additionally proposed system simplifies the idiom in terms of providing Gujarati meaning of that idiom. We believe that the output provided by the proposed system is the text without any Gujarati idiom will be further useful for the translation of Gujarati idiomatic text to any other language in the world.

References

[1] Wikipedia, "Gujarati language", https://en.wikipedia.org/wiki/Gujarati_language (accessed January 24, 2022).

[2] Wikipedia, "Gujarati grammar", https://en.wikipedia.org/wiki/Gujarati_grammar (accessed January 24, 2022).

[3] RitiRiwaz, "Gujarati Language Gujarati History and Facts", https://www.ritiriwaz.com/gujarati-language-gujarati-history-and-facts/ (accessed January 24, 2022).

[4] Wikipedia, "Stemming", https://en.wikipedia.org/wiki/Stemming (accessed January 24, 2022).

[5] GeeksforGeeks, "Introduction to Stemming", https://www.geeksforgeeks.org/introduction-to-stemming/ (accessed January 24, 2022).

[6] Bitext, "What is the difference between stemming and lemmatization?", https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/ (accessed January 24, 2022).

[7] Bing Microsoft Translator, "Microsoft Bing", Microsoft Corporation Ltd.; Available Online: https://www.bing.com/translator (accessed January 24, 2022).

[8] Wikipedia, "Microsoft Translator", Available Online: https://en.wikipedia.org/wiki/Microsoft_Translator (accessed January 24, 2022).

[9] Google Translate, "Google Translate", Google Corporation Ltd.; Available Online: https://translate.google.co.in/ (accessed January 24, 2022).

[10] Wikipedia, "Google Translate", Available Online: https://en.wikipedia.org/wiki/Google_Translate (accessed January 24, 2022).

[11] Saini J. R. and Modh J. C., "GIdTra: A dictionary-based MTS for translating Gujarati bigram idioms to English," 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2016, pp. 192-196, doi: 10.1109/PDGC.2016.7913143.

[12] 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.

[13] Modh J. C. and Saini J. R., "Context Based MTS for Translating Gujarati Trigram and Bigram Idioms to English," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-6, doi: 10.1109/INCET49848.2020.9154112.

[14] 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. http://dx.doi.org/10.14569/IJACSA.2021.0120128.

[15] Modh J. C. and Saini J. R., "Dynamic Phrase Generation for Detection of Idioms of Gujarati Language using Diacritics and Suffix-based Rules", International Journal of Advanced Computer Science and Applications(IJACSA), 12(7), 2021. http://dx.doi.org/10.14569/IJACSA. 2021.0120728.

[16] Grace Muzny and Luke Zettlemoyer, "Automatic Idiom Identification in Wiktionary", The Stanford NLP Group; Available online: https://nlp.stanford.edu/~muzny/docs/mz-emnlp2013.pdf.

[17] Rakesh Verma and Vasanthi Vuppuluri, "A New Approach for Idiom Identification Using Meanings and the Web", Available online: http://www2.cs.uh.edu/~rmverma/ranlp.pdf.

[18] Ferdy Hubers, Catia Cucchiarini, Nicoline van der Sijs, "Knowledge of idiomatic expressions in the native language: Do emigrants lose their touch?", ScienceDirect, 2022, Available online: https://www.sciencedirect.com/science/article/pii/S0024384122000031.

[19] Ruth Kessler & Claudia K. Friedrich, "Delayed prediction of idiom constituent meaning points to weak holistic multi-word representation in children, Language, Cognition and Neuroscience", 2022, DOI: 10.1080/23273798.2022.2035781.

[20] Bakir and Ricky Umbu, "An Analyze Typology of Idiomatic Expression in American Pie Movie", 2022, Available online: https://eprints.umm.ac.id/83454/.

[21] Muhammad Azimi Ramadhan, Supiani, Angga Taufan Dayu, "The Adoption of English Movies in Learning English Idioms: The English Undergraduate Students' Perception of English Movies with English Subtitles", 2022, Available online: http://eprints.uniska-bjm.ac.id/9595.

[22] GujaratiLexicon, Gujaratilexicon.com, Available online: http://www.letslearngujarati.com/about-us (accessed January 24, 2022).

[23] Rudhiprayog ane kahevatsangrah, published by Director of Languages, Gujarat State, Gandhinagar. 2010.