Using IndoWordNet for Contextually Improved Machine Translation of Gujarati Idioms

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 language is the Indo-Aryan language spoken by the Gujaratis, the people of the state of Gujarat of India. Gujarati is the one of the 22 official languages recognized by the Indian government. Gujarati script was adopted from Devanagari script. Approximately 3000 idioms are available in Gujarati language. Machine translation of any idiom is the challenging task because contextual information is important for the translation of a particular idiom. For the translation of Gujarati idioms into English or any other language, surrounding contextual words are considered for the translation of specific idiom in the case of ambiguity of the meaning of idiom. This paper experiments the IndoWordNet for Gujarati language for getting synonyms of surrounding contextual words. This paper uses n-gram model and experiments various window sizes surrounding the particular idiom as well as role of stop-words for correct context identification. The paper demonstrates the usefulness of context window in case of ambiguity in the meaning identification of idioms with multiple meanings. The results of this research could be consumed by any destination-independent machine translation system for Gujarati language.**

***Keywords*—Contextual information; Gujarati; idiom; IndoWordNet; Machine Translation System (MTS); n-gram model**

I. Introduction

This Machine Translation (MT) is the application of Natural Language Processing (NLP) which is an area of Artificial Intelligence (AI). Machine Translation is the need for the communication between people knowing two different languages. Gujarati language has more than 46 million speakers worldwide making it the 26th spoken native language in the world [1].

Idiom is a common phrase whose meaning is different from its individual literal meaning of word. It is widely used and it has its popular meaning. Gujarati language has approximately 3000 n-gram idioms. Meaning of Gujarati idiom can be understood by the context of the text. Here context refers to the information surrounding that idiom which helps in understanding the meaning of idiom. Dictionary based approach can be used for single meaning idiom unless it has more than one possible meaning. For multiple meaning idioms, the context information before and after the Gujarati idiom appearing in the text has to be looked. Contextual information is nothing but the words surrounding the specific idiom used in the text.

A. IndoWordNet

IndoWordNet is large linked lexical database for Indian languages including Gujarati language. IndoWordNet is the WordNet for Indian languages developed by Center for Indian Language Technology (CFILT) in the Computer Science and Engineering Department at IIT Bombay. Nouns, adjectives, verbs and adverbs are grouped into set. Gujarati WordNet is very important resource for the natural language processing task [2-5].

B. Gujarati Stop-Words

The Stop-words are the most common words in the particular language. They do not add meaning to the text. For natural language processing task, stop-words are generally removed or ignored as pre-processing activity. For phrase searching, stop-words cannot be ignored [6]. Stop-words list is not common for all domains. Example of Gujarati stop-words are અથવા અને આ આથી આવે એ કે કોઈ છતાં છે છો જ જેમ જો તેમ પછી પણ માટે હોય etc. [7].

C. N-gram

N-gram is a contiguous sequence of n items from a given text [8]. N-gram of size 1 is known as 1-gram or unigram; size 2 is referred as 2-gram or bigram; size 3 is referred as 3-gram or trigram; size 4 is referred as 4-gram or four-gram and so on. If input text is “I love my country”, then examples of bigrams are “I love”, “love my” and” my country”; examples of trigrams are “I love my” and “love my country”. N-gram model is used in natural language processing. 1-gram to 8-gram generation sequence will generate first 1-gram, then 2-gram,…8-gram; whereas 8-gram to 1-gram generation sequence will generate first 8-gram, 7-gram, 6-gram,…1-gram respectively.

The rest of the paper is organized as follows: Section II presents the literature review related to context and idiom translation; Section III covers the methodology including idiom data collection and proposed algorithm to find the meaning of idiom. In Section IV, extensive experiments with results and analysis are discussed using IndoWordNet and contextual information; finally conclusion, limitation and future direction are described in Section V.

*Corresponding Author

II. Related Literature Review

For the machine translation from one language to other language, several projects have been carried out. For the Machine Translation from Gujarati to English language, Google and Microsoft are the big players in the market. Google Translate [9] supports more than 100 languages, while the Microsoft Translator [10] supports 54 languages. Both support translation from Gujarati to English language. Both do literal translation for Gujarati idioms. Context identification is very important for translation of idioms. Various work related to context identification and idiom translation carried out.

Fortu et al. [11] proposed algorithm for detecting context boundaries and used machine learning model for the detection of subjective contexts using a set of syntactic features. They categorized various types of contexts like Subjective, Time/Space, Domain, Necessity, Planning/Wish contexts.

Turney [12] defined feature relevance definitions like strongly relevance and weekly relevance. He defined various context related definitions like primary feature, contextual feature, context-sensitive feature, strongly context-sensitive features and illustrates these definitions.

Leacock et al. [13] proposed statistical classifier for the identification of word sense. Their proposed classifier is used to disambiguate adjective, verb, and nouns. They combined local clues with topical context. They used general text corpus for training examples. They concluded that the local context is superior to topical context.

Mishra et al. [14] designed hybrid approach to automate Hindi to English idiom translation. They collected idioms in the form of Hindi-English language pair and classified idioms in three categories: (i) similar meaning and similar form (ii) similar meaning and dissimilar form (iii) different meaning and different forms in both languages. They used transfer-based and interlingual-based machine translation of rule based approach.

Pedersen et al. [15] used SenseClusters [16], freely available intelligent system that clusters similar context texts in natural language text. SenseClusters is purely unsupervised and language independent approach. SenseClusters system supports different context representation schemes, feature selection from large corpora, various cluster algorithms and labels for clusters.

Sekiya et al. [17] used Reuters news articles and focused on determining all the senses for every word. They generated conceptual fuzzy sets to express word senses and five statistical measures as relations. They calculated cogency and mutual information by comparing compatibility between each measure and prediction model. They demonstrated the usefulness of the word sequences to identify context. They focused just four words before the target word in experiments.

Salton et al. [18] applied substitution based technique for English/Brazilian-Portuguese language pair. They first substituted original idiom with its literal meaning before translation and again substituted literal meaning with idioms following translation. They indicated improved performance.

Based on this literature review of the most relevant research works found in research community and the analysis based on context identification and Gujarati idiom translation, no researchers have done context identification for Gujarati language idioms. No researchers have experimented window sizes on Gujarati idioms for correct meaning identification. Most of the researchers have applied various techniques for determining word sense; some researchers have applied idiom translation techniques other than Gujarati language.

III. Methodology

A. Data Collection

Gujarati language idioms are collected from different 11 books and websites. Idioms can be classified as bigram, trigram, four-gram, five-gram, six-gram, seven-gram, eight-gram and so on. Out of 2908 idioms, 1735 idioms are bigrams and 892 idioms are trigrams. Total bigram and trigram idioms are 2627. So 90% of total idioms are bigrams and trigrams. Only 281 idioms are from other category like monogram, four-gram, five-gram, six-gram, seven-gram, eight-gram and so on. So, the analysis of bigram and trigram idioms was done first. Table I shows the classification of Gujarati Idioms. It is based on the work of Modh and Saini [22].

Idioms can be classified further on the base of its meanings like 1-meaning, 2-meanings, 3-meanings, 4-meanings and so on. For example "સંસાર માંડવો" 'sansar mandvo' is a Gujarati bigram single-meaning i.e. 1-meaning idiom and its meaning in Gujarati is "પરણવું" 'paranavu' only and its translation in English language is "to marry"; where as "આંખ બતાવવી" 'aankh batavavi' is a Gujarati bigram 2-meaning idiom because it has two possible meanings in Gujarati as "ધમકી આપવી" 'dhamaki aapvi' and "આંખ બતાવવી" 'ankh batavavi' and so two corresponding possible translations in English language are "to threaten" and "show eyes". In the collection of overall 2627 bigram and trigram idioms, it was found total 2455 single meaning idioms and 172 idioms are having more than 1-meaning. From bigram and trigram idioms, 172 idioms are having 2-meaning, 3-meaning and 4-meaning idioms [19]. Table II shows the classification of bigram and trigram idioms on the base of meanings of idioms. It is based on the work of Modh and Saini [22].

TABLE I. Classification of Gujarati idioms

N-gram Idiom Category	Count
Bigrams (n=2)	1735
Trigrams (n=3)	892
Other N-Gram idioms where n>=4	281
Total	2908

TABLE II. Classification of Gujarati Bigram and Trigram idioms on the base of Meanings

Meanings	Bigram Count	Trigram count	Total Count
1-meaning (single-meaning)	1675	780	2455
n-Meanings where n>=2	60	112	172
Total	1735	892	2627

If idiom has single meaning, then English translation of that particular idiom is very simple and direct, algorithm has to replace its meaning in the place of that idiom. If the idiom has more than one possible meaning, then contextual information comes in the picture. Contextual information is nothing but the collection and study of surrounding words before and/or after the particular idiom. For the correct translation of particular idiom having multiple meanings, algorithm has to examine the surrounding words before and/or after the particular idiom. By removing stop-words from the surrounding words, contextual words are obtained. Fig. 1 and Fig. 2 show the graphical representation of contextual words with bigram and trigram idiom respectively.

So here three options can be considered for contextual words; 1) contextual words before idiom i.e. left window only 2) contextual words after idiom i.e. right window only 3) contextual words before and after idiom i.e. left window and right window collectively. One more concern is about how many surrounding words to be verified from the given input text for the precise meaning identification of particular idiom. Three cases were experimented and results were recorded in order to identify the correct window size for left, right, both and optimum window size for the translation of Gujarati idiom(s) from the given Gujarati input text.

B. Software and Tools used

Following is the list of software and tools that are used to implement the proposed methodology.

- Spyder 4.1.5 (Scientific Python Development Environment) IDE.
- Anaconda3 2019.03 (Python 3.7.3 64-bit).
- pyiwn (Python-based API for IndoWordNet).
- Windows 10 (Operating System).
- XAMP 7.4.11 (cross-platform local web server).
- MySQL (database to store idioms).
- PHP 7.4.11 (scripting language for web development).
- Sublime Text & Visual Studio Code (editors).

C. Algorithm

Table III shows the partial database of Idioms stored in Idiom table. In the database, only bigram and trigram idioms having more than one-meaning are shown. Researchers had already experimented with single meaning idioms [20-22]. “Idiom” field stores the bigram/ trigram/n-gram idiom. “Gujarati meaning” field stores meaning of particular idiom in Gujarati language. “English meaning” field stores the translation of particular Gujarati idiom in English language. “Gujarati Context Words” field stores the Gujarati context words related to particular idiom record. Gujarati Context Words are collection of all words from manually collected contextual words (from the corpus related to meaning of that idiom) and generated synonyms using Gujarati WordNet. If particular idiom has single meaning then only single record is there in the database. If idiom has n meanings, then n record entries are there in the database. For example, આંખ બતાવવી ‘aankh batavavi’ idiom has two possible meanings in Gujarati language, so two possible translations in English language; ધમકી આપવી ‘dhamaki aapvi’ (to threaten) and આંખ બતાવવી ‘aankh batavavi’ (literally show eyes e.g. for medical checkup). If this idiom has been used in given text, then algorithm has to decide any one meaning from the two possible meanings. “Gujarati Context Words” fieldis used for the context identification of particular idiom. If surrounding contextual words are related to ડોક્ટર તકલીફ દવા ઝાંખપ દેખાવું દૂર નજીક આંખ નજર વાંચન સમસ્યા then the translation of idiom આંખ બતાવવી ‘aankh batavavi’ is “show eyes” in English language and “આંખ બતાવવી” in Gujarati language. If surrounding words are related to લડાઈ બાળક ઠપકો માબાપ સજા છોકરો then the meaning of idiom આંખબતાવવી ‘aankh batavavi’ is “to threaten” in English language and “ધમકીઆપવી” in Gujarati language. Gujarati WordNet i.e. IndoWordNet was used for the collection of more contextual words on the base of synonyms of manually collected contextual words. Some words are not found in the Gujarati WordNet. Thus those more words were added in the field “Gujarati Context Words”. For example, words like કકળાટ ‘kakalaat’, રાજકારણ ‘raajkaaran’, સરોગેટ ‘saroget’, ચોમાસું ‘chomaasun’, ઝગડો ‘zhagado’ વોટીંગ ‘voting’, મધર ‘madhar’ etc. are frequently used words in Gujarati text but are not available in Gujarati WordNet. So these words were added in corresponding field of “Gujarati Context Words”. Gujarati Context Words play very important role in deciding the meaning of particular idiom and so the translation of Gujarati idioms. Algorithm calculated the frequency count of surrounding contextual words for each possible meaning of particular idiom by comparing with “Gujarati Context Words” column; the more count of context words decide the particular meaning of the idiom. “Popularity” field decides the more frequent meaning of the particular idiom assigned by the Gujarati expert(s) in case of ambiguity. For example, if particular idiom has 3 possible meanings, popularity value 1 is given to that record which meaning is more frequently used in real life. The particular record was decided by studying real life examples as well as with the help of Gujarati language expert(s). Only when there occurs a tie during the process of selection of meanings, the algorithm use “Popularity” field.

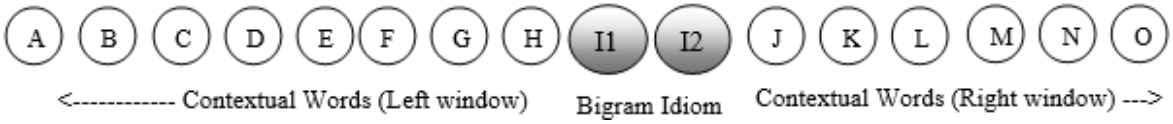

Fig. 1. Graphical Representation of Bigram Idiom with its Contextual Words.

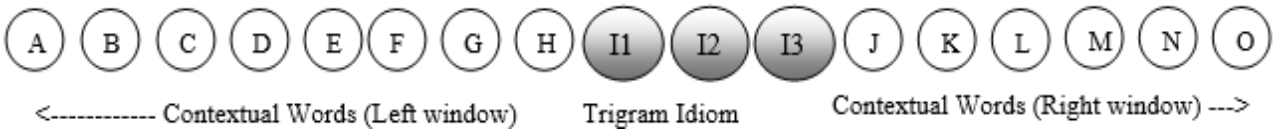

Fig. 2. Graphical Representation of Trigram Idiom with its Contextual Words.

TABLE III. IDIOM DATABASE (PARTIAL) FOR BIGRAM AND TRIGRAM IDIOMS HAVING MORE THAN ONE MEANING

Sr no (1)	Idiom (2)	Transliteration of (2) (3)	Gujarati Meaning (4)	English Meaning (5)	Gujarati Context Words (6)	Popularity (7)
1	આંખ બતાવવી	Aankh batavavi	ધમકી આપવી	To threaten	અંબા અથડામણ અનુશય અપકોશ અપચાર અબ્બા અબ્બાજાન અભિગ્રહ અમ્મા અમ્માજી અમ્મી અર્થદંડ અર્ભ અસ્મૃતિ આત્મજ આત્મસંભવ આયોધન આળ આસ્કંદ આહર આહવ ઉપાલંભ ઉલાહના કંકાસ કજિયો કતલ કલહ કલેશ કસૂર કાપાકાપ કિબલા કિબ્લા કિશોર કુમાર ક્ષતિ ખટરાગ ખતા ખામી ખિજાવું ખૂન ખૂનરેજી ખૂનામરકી ગફલત ગિલ્લા ઘર્ષણ ઘસારો ઘાત ચકમક ચિરંજીવ ચૂક છૈયો છોકરાં છોકરૂં છોકરો છોરો જંગ જનક જનની જનેતા જન્મદાતા જન્મદાત્રી ઝંઝટ ઝગડો ઝઘડો ઝપાઝપી ઝાટકણી ટંટો ઠપકારવું ઠપકો ડાંટ તકરાર તનય તનૂજ તાત દંડ દિકરો દ્વંદ્વ દ્વન્દ્વ ધમકી ધાત્રી ધિગાણું ધિક્કાર ધુત્કાર નંદન નવજાત શિશુ નવજાતક નાનકુ નિંદા પંચાત પિતા પિતાજી પુત્ર પુષ્કર પૂત પૃથુક પ્રતિદારણ પ્રહરણ ફટકાર ફરજંદ ફિટકાર બખેડો બચ્ચા બચ્ચું બટુક બાપ બાપા બાલક બાળ બાળક બાળકને બાળકો બેટો ભર્ત્સના ભાંડવું ભૂલ ભૂલકણાપણું મા માં માતા માતારી માતૃ માતૃકા માથાકૂટ માદર માબાપ માયા મારવું મારામારી મૂઠભેડ મૈયા યુદ્ધ રકઝક રણ લડકો લડવું લડાઇ લાડલો લાલ લોયો વટુ વઢવું વત્સ વધ વાદવિવાદ વાલિદ વિગ્રહ વિરોધ વિવાદ વિસ્મરણ વિસ્મૃતિ શિકાયત શિક્ષા શિફા શિશુ સંકુલ સંગ્રામ સંઘર્ષ સંતાન સજા સમર સામનો સુત સ્કંધ સ્મૃતિવિહીનતા હત્યા હિંસન હિંસા હુમલો	2
2	આંખ બતાવવી	Aankh batavavi	આંખ બતાવવી	show eyes	અંકુરણ અંકુરણ બિંદુ અંતર અંતરવેદના અંબક અક અધ અડચણ અધ્યયન અભ્યાસ અરિષ્ટ અર્વાક અલગ અવસન્નતા અવસન્નત્વ અવિદૂર અશર્મ અસુખ અસુવિધા આંખ આંખો આધું આધે આદીનવ આપત્તિ આપદ આપદા આફત આભીલ આર્તિ ઈક્ષણ ઈક્ષિકા ઉતાપો ઉદ્વેગ ઉપચારક ઉપતાપ એક તરફ એક બાજુ ઓળખ ઓસડ ઔષધ ઔષધિ કને કષ્ટ કુદૃષ્ટિ ક્લેશ ખ્યાલ ગભરામણ ચક્ષુ ચશ્મ ચિકિત્સક ચિતવન છેટું છેટે જડીબુટ્ટી જીવ પડવો જોડે ઝળકવું ઝાંખપ ઝાંખું ડીઠ ડૉક્ટર ડૉક્ટરને ડોક્ટર ડોક્ટરને તકલીફ તપાસ તસ્દી તેવર તોદન દરદ દરમાન દર્દ દર્શન થવા દવા દવાદારુ દાક્તર દારુ દુ દુખ દૂર દૂરત્વ દૂરવર્તી દૂરસ્થ દૂરસ્થિત દૃષ્ટિ દૃષ્ટિ પડવી દૃષ્ટિકોણ દેખરેખ દેખાવું દોયન ધૂંધલાઇ ધૂંધલાપણ ધૂંધળાઈ ધૂંધળાપણું ધ્યાન નજર નજર પડવી નજરિયા નજીક નયન નિકટ નિગાહ નિરીક્ષણ નેણ નેત્ર નેત્ર-દૃષ્ટિ નેન પઠન પડખે પડદો પડવું પરખ પરિપ્રેક્ષણ પરે પરેશાની પાથિ પાર્શ્વ પાસે પિઠ પીડા ફણગો ફાસલો બગલ બાજુ ભવાં ભૈષજ્ય મનોવ્યથા મુશ્કેલી મુસીબત યંત્રણા યાદ રસાયન રોહજ ર્દગ લોયન વાંયન વિપત્તિ વિપદ વિપદા વૃજિન વેગળું વેગળે વેદના વૈદ્ય વ્યથા વ્યાકુલતા વ્યાકુલપણું વ્યાકુળતા વ્યાધિ શૂળ સંકટ સંતાપ સંભાળ સમસ્યા સમીપ સુધ સુધિ સોય સ્મૃતિ હકીમ હૂક હેરાનગત	1
3	આંખ માંપાણી આવવું	Aankhma pani avavu	આંખમાં પાણી આવવા	water in the eye	અંકુરણ અંકુરણ બિંદુ અંબક અભિયોગ આંખ આજાર આતપ આમય આરજા ઈક્ષણ ઈક્ષિકા ઇલાજ ઉધરસ ઉપઘાત ઉપચાર કાશ કાસ કેસ ખટલો ખાંસી ગેસ યક્કર યક્ષ્મ યશ્મ ચિકિત્સા જર જુખામ જ્વર ટાઢ ઠંડક ઠંડી ઠસક ઠાંસો તખ્તો તરિયો તાપ તાવ ત્રિપાદ દરદ દવા-દાવા દારુ-પીનસ પાથિ નેન નેત્ર નેણ નયન દૂ દુખાવો દાવો અરજી ફણગો બિમારી બીમારી બુખાર માથું માવજત મુકદ્દમો મુકદ્દમો મુકદ્દમો રોગ રોગોપચાર રોહજ ર્દગ લક્ષણ લોયન વિકાર વૈદ્યકી વ્યાધિ શરદી શીત સંભાળગત સરદી સળેખમ સારવાર સારસંભાળ સૂકી ખાંસી હકીમી	1
4	આંખ માં પાણી આવવું	Aankhma pani avavu	દયાની લાગણી થવી	feeling compassion or affection	અંતરવેદના અક અક્રિયન અધ અડચણ અનસ્તિત્વ અનુપલબ્ધિ અનૈશ્વર્ય અપૂર્ણતા અપ્રાપ્તિ અભવ અભાવ અરિષ્ટ અલ્પતા અલ્પત્વ અવસન્નતા અવસન્નત્વ અવસ્થા અશર્મ અસંપન્ન અસમૃદ્ધ અસુખ અસુવિધા આદીનવ આપત્તિ આપદ આપદા આફત આભીલ આર્તિ આલમ ઉતાપો ઉદ્વેગ ઉપતાપ કંગાલ કંગાલિયત કકળાટ કમી કષ્ટ ક્લેશ ખામી ખાલીપણું ખોટ ગતિ ગભરામણ ગરીબ ગરીબાઈ ગરીબી ગેરહાજરી ઠાલાપણું તંગહાલ તકલીફ તસ્દી તોદન દરદ દરિદ્ર દરિદ્રતા દરિદ્રાણ દર્દ દળદર દશા દારિદ્ર દારિદ્રય દીન દીનતા દીનહીન દુ દુખ દૈન્ય દોયન ધનહીન નિધની નિર્ધન નિર્ધનતા પરેશાની પિઠ પીડા ફકીરી બિયારું બેહાલ મનોવ્યથા મુફલિસ મુશ્કેલી મુસીબત યંત્રણા રંક રંકતા રહિતપણું રાંક રાહિત્ય રિક્તતા લાઘવ વિધન વિધનતા વિપત્તિ વિપદ વિપદા વિપન્નતા વૃજિન વેદના વ્યથા વ્યાકુલતા વ્યાકુલપણું વ્યાકુળતા વ્યાધિ શૂન્યતા શૂળ સંકટ સંતાપ સૂરત સ્થિતિ હાલ હાલત હૂક હેરાનગત	2
5	આંખ માં પાણી આવવું	Aankhma pani avavu	આંખમા આંસુ આવવા	tears in the eyes	અર્ભ અશ્રુપાત આક્રંદ આક્રંદ કરવો આક્રંદન આત્મજ આત્મસંભવ કલ્પાંત કિશોર કુમાર ક્રંદન ચિરંજીવ છૈયો છોકરૂં છોકરો છોરો તનય તનૂજ દાદા દિકરો નંદન નવજાત શિશુ નવજાતક નાનકુ નાના પુત્ર પૂત પૃથુક ફરજંદ બચ્ચા બચ્ચું બટુક બાલક બાળ બાળક બેટો માતામહ રડવું રુદન રુદન કરવું રોદણું રોવું લડકો લાડલો લાલ વટુ વત્સ વિલપન વિલાપ વિલાપ કરવો શિશુ સુત	3

Input text is given in Gujarati language. Input text may contain idiom(s). Entire input text is searched for the idiom(s) using n-gram model. If idiom(s) found in the text, then it may be single meaning or it may be more than one meaning idiom.

For single meaning idiom, the "Gujarati meaning" or "English meaning" column of that idiom can directly be used. But if the idiom has more than one meaning, algorithm has to consider "Gujarati Context Words" column. The algorithm decides the meaning of the particular idiom and substitutes the particular idiom with "Gujarati meaning" column value and produce intermediate output in Gujarati language itself. Output contains Gujarati literal text without any idiom. The algorithm can generate n-gram from the given input text using both the sequence 1-gram to 8-gram or 8-gram to 1-gram.

In the next section, empirical results are shown.

IV. Results and Analysis

A. Experiments

For the experiments, 150 different Gujarati texts containing 30 different Gujarati idioms having single/multiple meanings from the various Gujarati websites as well as from offline Gujarati content were collected. The collection of various idioms within input texts was performed; like single idiom with single meaning, single idiom with more than one meaning(s), two idioms with single/multiple meaning(s), three idioms with single/multiple meaning(s), four idioms with single/multiple meaning(s) and so on.

L notation for Left window and R notation for Right window were used for simplification. (Ln, Rn) specifies Left window size n, Right window size n; (Ln, R0) specifies Left window size n and Right window size 0; (L0, Rn) specifies Left window size 0 and Right window size n; For example, Fig. 3 shows representation of (L6,R3). (L6, R3) denotes 6 words left side of the idiom and 3 words right side of the idiom. Surrounding words may or may not provide contextual information. Stop-words should be removed to get only contextual words information.

1) Experiment-1: Experiement-1 was conducted to decide two things (a) importance of various windows left, right or both for context identification (b) N-gram generation from the input text is possible by two ways; 1-gram to 8-gram generation sequence and 8-gram to 1-gram generation sequence. 1 to 8 gram generation sequence will generate first 1-gram, then 2-gram, 3-gram,…8-gram. 8 to 1 gram generation sequence will generate first 8-gram, then 7-gram, 6-gram,…1-gram. Which sequence is to be selected for better results? 150 Gujarati input texts containing single idiom only for each text was experimented. Idiom within text may have single/multiple meaning(s). Three cases were experimented (1) left window only (Ln,R0) (2) right window only (L0,Rn) and (3) left and right window both (Ln,Rn). Experiments for both the sequences for N-gram generation were conducted: 1-gram to 8-gram and 8-gram to 1-gram generation sequence. The algorithm will generate both the sequences by selection. For simplification and for evaluating importance of windows (left/right/left-right), all surrounding words of idioms were considered as contextual words and for that window size n=30 was applied for the experiment.

- Case-1: Using left window only for contextual information. The left window size was fixed as 30 and right window size was fixed as 0. Overall 150 input texts were tested for (L30,R0). Out of 150 input texts, idioms meaning precisely identified from 111 texts with 1 to 8 gram generation sequence (74% accuracy); idioms meaning precisely identified from 117 texts with 8 to 1 gram generation sequence (78% accuracy).
- Case-2: Using Right window only for contextual information. The right window size was fixed as 30 and left window size fixed as 0. Overall 150 input texts were tested for (L0,R30). Out of 150 input texts, idioms meaning precisely identified from 93 texts with 1 to 8 gram generation sequence (62% accuracy); idioms meaning precisely identified from 99 texts with 8 to 1 gram generation sequence (66% accuracy).
- Case-3: Using fixed left and fixed right window for contextual information. The left widow size was set as 30 and right window size 30 i.e. (L30,R30) and tested 150 input texts; Out of 150 input texts, idioms meaning precisely identified from 132 texts with 1 to 8 gram generation sequence (88% accuracy); idioms meaning precisely identified from 138 texts with 8 to 1 gram generation sequence (92% accuracy). Idioms meaning can't be identified from 12 input texts. These 12 input texts were examined and found that the total words in these 10 input texts were less than or equal to 10. Hence out of these 12 input texts, 10 texts have not sufficient contextual information before and/or after idiom.

By comparing Case-1 (left window only), Case-2 (right window only) and Case-3 (left and right window) results of Table IV, Case-3 results are clearly front runner. So it is concluded that only left window or only right window is not useful at all for identifying contextual information. Case-3 (both the left window and right window) i.e. context words before and after the idiom must be considered for collecting contextual information. Also got better results in 8-gram to 1-gram generation sequence compared to 1-gram to 8-gram generation sequence; Particular cases were observed and found that intermediate translation of particular Gujarati idiom will also generate Gujarati idiom; this generated idiom can be found with 8-gram to 1-gram generation sequence. For example, એક ઘાએ બે કટકા થવા 'ek ghaae be katkaa thavaa' is 5-gram idiom and its meaning is તડ ને ફડ જવાબ થવો 'tad ne fad javaab thavo', તડ ને ફડ 'tad ne fad' is 3-gram idiom and its meaning is સ્પષ્ટ 'spashta' or 'clear answer'; By taking sequence of 1-gram to 8-gram generation, તડ ને ફડ 'tad ne fad' idiom cannot be identified. So 8 to 1 gram generation sequence is preferred over 1 to 8 gram generation sequence.

Results of Experiment-1 in Table IV concluded two things (a) Left Window and Right Window both are required for contextual information. (b) For all N-gram idioms search within input text, 8-gram to 1-gram generation sequence is better one.

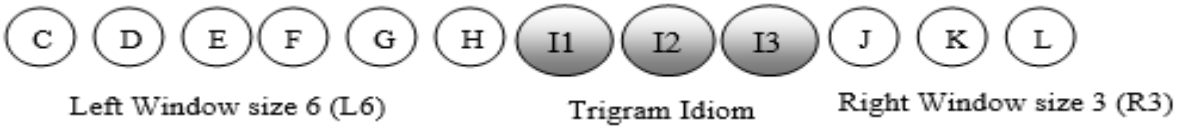

Fig. 3. Graphical Representation of (L6,R3).

TABLE IV. ACCURACY OF CORRECT MEANING IDENTIFICATION FOR SINGLE IDIOM HAVING SINGLE OR MULTIPLE MEANING (S)

Contextual Information and N-gram generation sequence	**Case-1 Left Window only (L30, R0)**	**Case-2 Right Window only (L0,R30)**	**Case-3 Left & Right Window (L30,R30)**
with stop-words and 1-gram to 8-gram generation	74%	62%	88%
with stop-words and 8-gram to 1-gram generation	78%	66%	92%

2) Experiment-2: Applying these two settings in the algorithm, experiment-2 was performed in which three things were evaluated: (1) different left and right window sizes for context identification (2) inclusion of stop-words or removing stop-words as contextual information (3) Using Gujarati WordNet words only as contextual information or with added manually collected words in WordNet words as contextual information. Two databases were used, in which first database contains only contextual words supported by Gujarati WordNet as "Gujarati Context Words" column and second database contains WordNet Words of first database + added contextual words in "Gujarati Context Words" column. These added words are not available in IndoWordNet.

For experiment-2, input texts with the sufficient contextual information i.e. input texts with at least ten words surrounding idiom(s) were selected. 8-gram to 1-gram generation sequence was set as it provides better results as per the experiment-1 results. Left window size was set variably from 1 to 20 and right window size was set variably from 1 to 20. Overall 150 input texts containing 30 multiple meaning idiom(s) were experimented. For this experiment input texts containing more than one multiple meaning idiom(s) were selected. The same texts with the inclusion of stop-words as contextual information and without consideration of stop-words as contextual information were experimented. In other words, overall 150 Gujarati input texts for (L1,R1), (L2,R2), (L3,R3),……upto (L20,R20) were tested. Only feasible window size(s) were selected for the experiment.

Table V shows the Experiment-2 results. Accuracy was calculated on the base of number of idioms meanings correctly identified. Combination of "Without stop-words and with All words (WordNet+Added Words)" shows the better accuracy for meaning identification for multiple meanings idioms; for (L2,R2) it shows 66.67% accuracy; for (L4,R4) it shows 83.33% accuracy; while for (L7,R7), (L10,R10), (L15,R15) it shows 100% accuracy. In other words, it gives correct translation for (L7,R7) to (L10,R10) and even for (L15,R15); for bigger window sizes (L20,R20) it reduces the performance (83.33% accuracy). More window size is not preferable for meaning identification of multiple meaning idioms. Moreover Table V shows that "without stop-words" option is giving better accuracy than "With Stop-words" option for all windows sizes.

Experiment-2 results concluded three things (1) stop-words should not be considered as contextual information i.e. from the input text stop-words should be ignored (2) All words (WordNet words+Added Words) should be used as "Gujarati Context Words' field for idiom database. Only WordNet words are not giving better results (3) At least Left window size 7 and right window size 7 are required to identify contextual words for the idiom having more than one possible meanings.

TABLE V. ACCURACY OF CORRECT MEANING IDENTIFICATION FOR MORE THAN ONE IDIOM HAVING MULTIPLE MEANING WITH 8-GRAM TO 1-GRAM GENERATION SEQUENCE

Using 8-gram to 1-gram generation sequence	**(L2,R2)**	**(L4,R4)**	**(L7,R7)**	**(L10,R10)**	**(L15,R15)**	**(L20,R20)**
With stop-words and with WordNet words only	16.67	16.67	33.33	66.67	83.33	66.67
Without stop-words and with WordNet words only	33.33	50	83.33	83.33	83.33	50
With stop-words and with All Words (WordNet+Added Words)	16.67	16.67	50	83.33	83.33	66.67
Without stop-words and with All words (WordNet+Added Words)	66.67	83.33	100	100	100	83.33

If particular Gujarati idiom has more than one possible meaning, then for the translation of that Gujarati idiom into English language, sufficient contextual information is required. Contextual words before the particular idiom and contextual words after the particular idiom must be considered as contextual information for the identification of the precise meaning of multiple meaning idioms. Stop-words should be ignored when considering contextual words before and after multiple meanings idioms.

Gujarati Context Words play very important role in the context identification of multiple meaning idioms. By studying corpus of each multiple meaning idiom used in real life, Gujarati Context Words can be collected. Using Gujarati WordNet, more synonym words can be added in the collection of Gujarati Context Words. Additional words which are not supported by IndoWordNet can be added into the database for collection of Gujarati Context Words collection. The compiled collection of Gujarati Context Words is required source for context identification in the algorithm.

If input Gujarati text has idiom with multiple meanings and if Gujarati text contains overall less than 10 words or 0 context word before idiom or 0 context word after idiom, then the precise meaning identification of that particular idiom is difficult and challenging. As per the experiments, it is suggested that, for correct meaning identification of Gujarati multiple meaning idioms, at least seven contextual words before and seven contextual words after that particular idiom should be verified.

V. Conclusion

Based on the results received from the intermediate translations of Gujarati idioms into literal Gujarati text, it is advocated that the proposed machine translation system is promising and worth implementation in real world for the translation of Gujarati idioms. Google Translate and Microsoft Translator also do the literal word to word translation in case of Gujarati idioms. The proposed system can be implemented for translation of Gujarati idioms to any other language translation as it is language independent. Proposed algorithm substitutes idiom with the literal text that can be used for any other language translation from Gujarati language.

Gujarati synonyms were collected from IndoWordNet and from the initially collected context words. Gujarati WordNet provides all the forms of the words in terms of nouns, adjectives, verbs and adverbs. Sometimes it provides additional synonyms not related to idiom meaning; even then it provides better results in terms of contextual words for identifying contextual information. Idiom meaning identification is not possible if idiom used in odd or strange context. In Gujarati, many words adapted from English are used frequently and those words are not included in Gujarati WordNet. So extensive corpus related to multiple meaning idioms is to be examined and used for further improvement.

In future, authors will extend the context identification for the n-gram idioms where n>=9; variety of window sizes can be tried out in future; experiments of window size with idioms of any language can be done. In future, authors are planning to implement and experiment using lemmatization and stemmer.

References

[1] Wikipedia, “Gujarati language”, https://en.m.wikipedia.org/wiki/Gujarati_language (accessed December 24, 2019).

[2] WordNet, “A Lexical Database for English”, Princeton University, Available Online: https://wordnet.princeton.edu/ (accessed December 24, 2020).

[3] Wikipedia, “IndoWordNet”, https://en.wikipedia.org/wiki/IndoWordNet (accessed December 24, 2020).

[4] IndoWordNet, “A wordnet of Indian languages”, Center for Indian Language Technology, Indian Institute of Technology, Bombay, Available Online: http://www.cfilt.iitb.ac.in/indowordnet/(accessed December 24, 2020).

[5] Ritesh Panjwani, Diptesh Kanojia, Pushpak Bhattacharyya, 2018, “pyiwn: A Python-based API to access Indian Language WordNets”, Proceedings of the GWC 2018 The 9th Global WordNet Conference, Nanyang Technological University (NTU) Singapore,8-12 January 2018, Available online: http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/GWC2018_paper_40.pdf.

[6] Wikipedia, “Stop word”, https://en.wikipedia.org/wiki/Stop_word (accessed December 24, 2020).

[7] Quanteda/Stopwords, 2019, “A List 210 Gujarati Stop Words”, Available online: https://github.com/quanteda/stopwords/files/2719155/A.List.of.210.Gujarati.Stop.Words.txt(accessed December 24, 2020).

[8] Wikipedia, “n-gram”, https://en.wikipedia.org/wiki/N-gram (accessed December 24, 2020).

[9] Google Translate, Google Corporation Ltd.; Available Online: https://translate.google.co.in/ (accessed December 24, 2020).

[10] Microsoft Translator, Microsoft Limited, Available Online: https://www.bing.com/translator (accessed December 24, 2020).

[11] Ovidiu Fortu and Dan Moldovan, 2005, “Identification of Textual Contexts”, International and Interdisciplinary Conference on Modeling and Using Context; Available online: https://link.springer.com/chapter/10.1007/11508373_13.

[12] Peter Turney, 2002, “The Identification of Context-Sensitive Features: A Formal Definition of Context for Concept Learning”; Available online: https://arxiv.org/ftp/cs/papers/0212/0212038.pdf.

[13] Claudia Leacock, Martin Chodorow, George A. Millers, 1998, “Using Corpus Statistics and WordNet Relations for Sense Identification”, Association for Computational Linguistics, Available online: https://www.aclweb.org/anthology/J98-1006.pdf.

[14] Himani Mishra, Rajesh Kumar Chakrawarti, Dr. Pratosh Bansal, 2017, “A New Approach for Hindi to English Idiom Translation”, International Journal on Computer Science and Engineering (IJCSE), Vol. 9 No. 07, Jul 2017, Available online: http://www.enggjournals.com/ijcse/doc/IJCSE17-09-07-024.pdf.

[15] Ted Pedersen and Anagha Kulkarni, 2005, “Identifying Similar Words and Contexts in Natural Language Using SenseClusters”, Association for the Advancement of Artificial Intelligence (AAAI); Available online: https://www.aaai.org/Papers/AAAI/2005/ISD05-013.pdf.

[16] SenseClusters, 2005, “Cluster contexts and words based on contextual similarity”, Available online: http://senseclusters.sourceforge.net/ (accessed December 24, 2020).

[17] Hiroshi Sekiya, Takeshi Kondo, Makoto Hashimoto, Tomohiro Takagi, 2007, “Context representation using word sequences extracted from a news corpus”, ScienceDirect, International Journal of Approximate Reasoning, 45 (2007) 424-438, Available online: http://www.sciencedirect.com/science/article/pii/S0888613X06001125.

[18] Giancarlo D. Salton and Robert J. Ross and John D. Kelleher, 2014, “Evaluation of a Substitution Method for Idiom Transformation in Statistical Machine Translation”, Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 38-42, Gothenburg, Sweden, 26-27 April 2014, Association for Computational Linguistics,Available online: http://www.aclweb.org/anthology/W14-0806.pdf.

[19] Gujarati Lexicon, Gujaratilexicon.com, Available online: http://www.letslearngujarati.com/about-us (accessed December 25, 2020).

[20] 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.

[21] 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/.

[22] 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/.