Gujarati Language: Research Issues, Resources and Proposed Method on Word Sense Disambiguation

Tarjni Vyas, Amit Ganatra

Abstract*: *Gujarati Word Sense Disambiguation (WSD) is an exceptionally complex when it comes to Natural language handling because it needs to manage complexities found in a language. In this paper, the discussion has put forward about Guajarati language, Gujarati Wordnet and Gujarati word sense disambiguation. Accordingly, the deep learning approach is found to perform better in Gujarati WSD yet one of its weakness is the prerequisite of enormous information sources without which preparing is close to impossible. On the other hand, utilizes information sources to choose the meanings of words in a specific setting. Provided with that, deep learning approaches appear to be more suitable to manage word sense disambiguation; however, the process will always be challenging given the ambiguity of natural languages.

Keywords: Word Sense Disambiguation, Gujarati Language, Deep learning, Natural language processing, Lesk Algorithm, Wordnet.

I. INTRODUCTION

A. ORIGIN OF GUJARATI LANGUAGE

There are 7 main families of the languages of the world. Out of these families, one language family from Indian- European origin has come as Indian Arya (Indo–Aryan) family. It is believed that Indian-Arya language has started from tenth-eleventh century to this date. Gujarati language too has originated during this time, developed and have reached to the time of today. Some part of Gujarat was under the reign of Gurjars and that part was known by the name of Gujarat or Gurjar. It is considered that Gujarat word has relation with this history. And from it, the name of the language was given as Gujarati. Gujarati language vocabulary is influenced by so many languages such as Sanskrit, Prakrit, Apbhransh, Arbi, Farsi, Portuguese and English.

Revised Version Manuscript Received on 16 September, 2019.

***** Correspondence Author

Tarjni Vyas*, Department of Computer Science and Engineering (CSE), Institute of Technology,
Nirma University, Ahmedabad, India. Email: tarjni.vyas@nirmauni.ac.in

Amit Ganatra, Dean-Faculty of Technology and Engineering, Devang Patel Institute of Advance Technology & Research (DEPSTAR), Charotar University of Science and Technology. Email:amitganatra.ce@charusat.ac.in

B. GUJARATI GRAMMAR

There are 32 consonants and 8 vowels in Gujarati language.

No	Components	Number	Details
1	Consonants	32	Fig2
2	Vowels	8	Fig2
3	Tenses	3	Past,Present, Future
4	Vachans	2	Singular,Plural
5	Sentence Structure	3	Subject,object,Verb

Sentence structure is made in the order of Subject, Object and Verb in Gujarati language.

There are three Tenses in Gujarati Past Tense, Present Tense and Future Tense. There are two types of Vachans in Gujarati.Singular and Plural. Gujarati Language includes following cases.(Table 1)

- Nominative Case
- Objective Case
- Instrumental Case
- Dative Case
- Ablative Case
- Possessive Case
- Locative Case

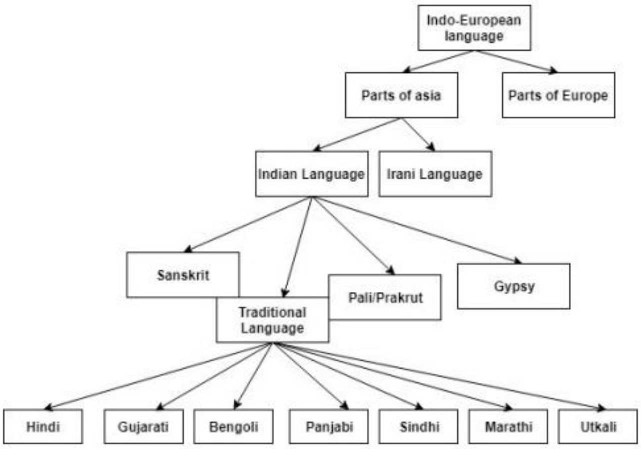


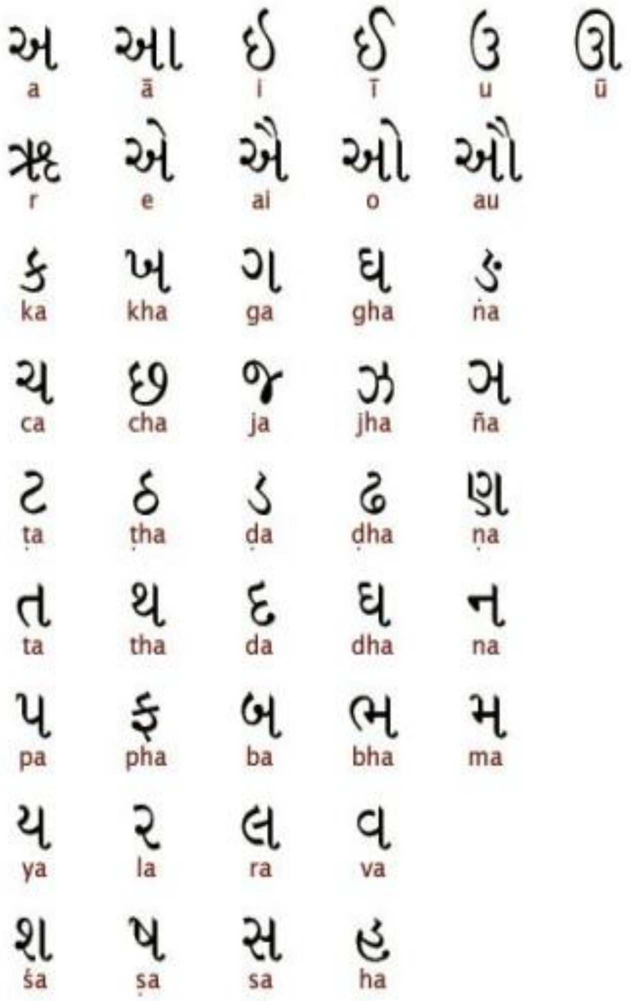
Figure 1 Gujarati Language Origin

Gujarati language is made up of letters which are called as "kakko" having consonants and vowels. following is the structure of Gujarati language where collection of letters make word and collection of words make term.

letter+letter+letter+ ... = word

word+word+word+... = term

term+term+term+... = Sentence



અ a	આ ā	ઇ i	ઈ ī	ઉ u	ઊ ū
ઋ r	એ e	ઐ ai	ઓ o	ઔ au	
ક ka	ખ kha	ગ ga	ઘ gha	ઙ ṅa	
ચ ca	છ cha	જ ja	ઝ jha	ઞ ña	
ટ ṭa	ઠ ṭha	ડ ḍa	ઢ ḍha	ણ ṇa	
ત ta	થ tha	દ da	ધ dha	ન na	
પ pa	ફ pha	બ ba	ભ bha	મ ma	
ય ya	ર la	લ ra	વ va		
શ śa	ષ ṣa	સ sa	હ ha		

Figure 2 Gujarati Language Letters

C. COMPLEXITIES IN GUJARATI LANGUAGE

- Gujarati Language is gender sensitive language. Every noun is having particular gender type from Feminine, Masculine and Neuter. same words with different genders are having different meaning.
- Word which has singular form may not have plural forms and words which has plural forms may not have singular forms.
- Gujarati language has varieties of dialects in the form of regional languages. Dialects depends on factors like region, society and community .In Gujarati language there are 4 main dialects.
    - Kathiyawadi Spoken Language
    - Madhya Gujarat/Charotar Spoken Language
    - South Gujarat or Surti Spoken Language
    - There are also many sub-dialects among these spoken languages.
- Gujarati language has words which are having multiple meanings.Those words are called as poly-semi words. poly-semi words often creates ambiguity in the statements. poly-semi words makes Gujarati language an ambiguous language.

II. LANGUAGE RESOURCES

WordNet is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets[1].

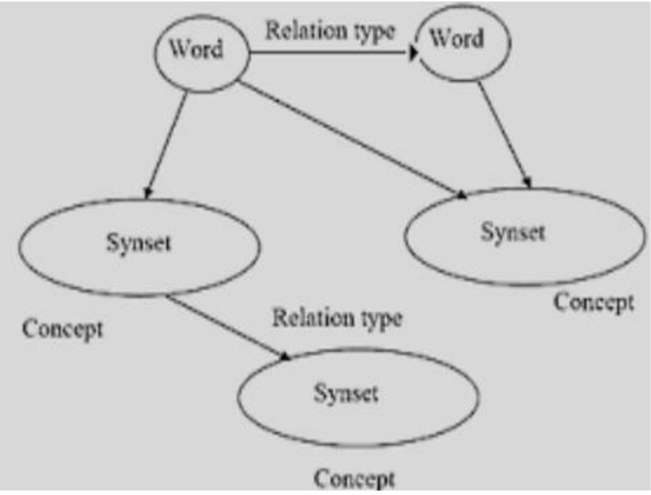


Figure 3 logical Structure of Wordnet [6]

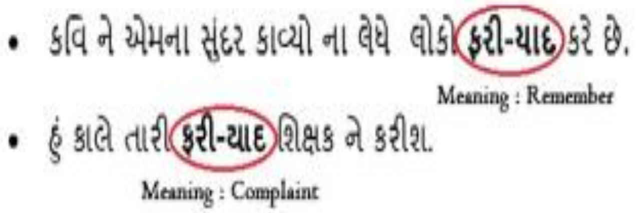
Gujarati Wordnet

Gujarati is one of the 22 official languages of India.[2] Gujarati Wordnet is being developed utilizing expansion approach with Hindi language.[2] Hindi Wordnet (Narayan D. et al., 2002) was the first wordnet for the Indian languages. Based on Hindi wordnet, wordnets for 17 different Indian languages are getting built using the expansion approach. One such effort is Gujarati wordnet[2]. Current Gujarati wordnet contains 35677 synsets.

Another online resource is called as Gujarati Wiktionary[11] which has different categories such as nouns, lemmas and others.

III. GENERAL WORD SENSE DISAMBIGUATION

One of the most prominent features of all the modern languages in use today is that they are inherently ambiguous. All the sentences and the word usages depend upon the context in conversation and hence when one tries to employ computational techniques especially during the study of natural language processing, a lot of confusion is bound to occur. The same is the case for sentences in Gujarati language. (Figure 4)



- કવિ ને એમના સુંદર કાવ્યો ના લેધે લોકો ફરી-યાદ કરે છે.

Meaning : Remember

- હું કાલે તારી ફરી-યાદ શિક્ષક ને કરીશ.

Meaning : Complaint

In above sentence it is seen that the same word means to "remember again" in the first sentence and then means “complaint” in the second sentence.

For a normal human to understand it is relatively very simple to grasp the context which is in the sentence but for any other applications like Machine based Translations, Text summarization applications standard method called word embeddings are used where each word in a given sentence corresponds to a vector in very high dimensional space.

Hence the word with the same characters would have the same embeddings and so when we train novel deep learning methods and recurrent neural networks upon these sentences then lead to confusion in the context of the sentence.

A. TRADITIONAL APPROACHES FOR GUJARATI WORD SENSE DISAMBIGUATION

Lesk algorithm works on theory of similarity between two words. Each word is having their unique definitions. Let’s say a word *"car"* can be defined as a vehicle with an engine which runs on a road. And a word *"boat"* can be defined as a vehicle with an engine which swims on a water. A third word *"sun"* can be defined as a huge sphere made of hydrogen and helium. Now if words car, boat and sun are to be compared by Lesk algorithm, car and boat are nearest to each other than sun. It happened because car and boat both has engine and both are vehicles but sun has neither of this.

There are number of variations of Lesk algorithm is available but two major type of Lesk algorithm are simplified Lesk algorithm Original Lesk algorithm and Adapted (or Extended) Lesk algorithm. Extended Lesk algorithm is given by Satanjeev Banerjee and Ted Pedersen, 2002/2003[4] Lesk algorithm operates on WordNet definitions. WordNet is a large database that contains definition of English words.

The original Lesk algorithm senses similarity of words between definitions of two words. Let’s take an example of comparison of two pairs car-boat and car-sun. When comparing car and boat: {vehicle, engine, runs, road} ∩ {vehicle, engine, swim, water} = {vehicle, engine}. The cardinality of resultant set is 2. In Lesk algorithm, cardinality can be taken as sense. When comparing car and sun: {vehicle, engine, runs, road}∩{sphere, hydrogen, helium} = Ø. The cardinality of resultant set is 0. It is clear that sense of car-boat is higher than sense of car-sun. So, Lesk algorithm outputs boat as nearest neighbor of car. This can help to solve disambiguate in natural language processing[4]. The simple Lesk algorithm demands comparison between exact word of definitions. To overcome limitations of simple Lesk algorithm, extended or adapted algorithm is proposed. In extended Lesk algorithm, a separate work vector is created for each and every word in WordNet database. A work vector contains similarity between words. A gloss vector is created for words and it contains similarity between work vectors of words in definition. Then these gloss vectors can be used to compare words using extended Lesk algorithm. Gloss vectors can be compare by using method of cosine similarity[4]

One word can have multiple definitions and presence of each and every word in definition matters. If a single word is missing in a definition then it changes relation between words radically. To overcome this problem of Lesk algorithm, different researches have given their contribution to this algorithm. They have used their own methods and subset of WordNet to improvise original and extended Lesk algorithm. Other variations of Lesk algorithm is given in timeline (Figure 5).

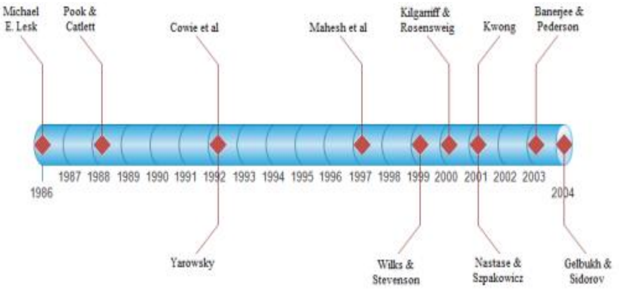


Figure 4 Timeline for Lesk algorithm

B. CURRENT APPROACHES AND ISSUES WITH GUJARATI WORD SENSE DISAMBIGUATION

The problem of word sense disambiguation is of a considerable age and has a firm footing in the NLP community, hence various algorithms and methods have been proposed Navigli in 2009 did comprehensive survey on the subject[6].The main task is to assign a specific word in the sentence a matching meaning in the WordNet which is a corpus of words and their corresponding meanings. Authors have also combined promising methods like the SVM (Support Vector Machines) to work upon specialized word embeddings. Various methods inspired from graph theory and graph traversal have been proposed.

In the recent years due to the rise in deep learning-based methods the application of LSTMs (Long Short-Term Memory) neural networks have considerably increased. The beauty of such approaches is the simplicity and the remarkable accuracy achieved and with the rise of the internet and rise of numerous data collection and arrival sources we are able to collect language data like never before in human history. One recently proposed LSTM architecture by Yeunn et al[10] trained on around 100 billion word corpus and with a fraction of sense embeddings were able to achieve the best known accuracy in WSD.

When one considers the problem of Word Sense disambiguation in other languages then the balls seems to roll out of the court as only a very small fraction of methods have been developed for European languages like German, French and Italian. For Indian Languages the progress in WSD has been near nil. The chief reason is the extreme scarcity of datasets in language other than English. In case of Indian languages there is almost no proper labeled language dataset. While by the work of P. Bhattacharya and others at IIT Bombay have proven to be a step-in right direction by making corpuses like Multilingual Wordnets possible.[7]

IV. DEEP LEARNING APPROACH WORD SENSE DISAMBIGUATION

Deep learning or machine learning approaches make use of frameworks that are prepared and skilled to deal with word sense disambiguation. Borah et al.[8] direct that in this technique, a classifier is designed and prepared, which is then used to assign meanings to concealed examples. In this methodology, the underlying input comprises of the words to be disambiguated, alongside content in which it is installed -

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

which is referred to as its setting.

Accordingly, this underlying input is prepared to utilize grammatical feature labelling or any morphological handling. After this initial preparing, Nameh[9] states that a fixed arrangement of linguistic highlights is extricated applicable to the learning task. These highlights can be either of two classes: co-occurrence or collocation[10] At the onset, co-occurrence highlights comprise of information about neighboring words. In this methodology, words themselves fill in as highlights. The value of a highlight is the occasions the word happens in the region encompassing the objective word. On the other hand, collocation highlights encode data about expressions of explicit positions that are situated to left or right of the objective word[10]. This way, typical highlights incorporate the word, the root type of word, together with the word's grammatical form.

V. PROPOSED MODEL FOR GUJARATI LANGUAGE WORD SENSE DISAMBIGUATION

Database:	Gujarati WordNet (35677 Synsets) and Wikipedia (27,800 pages)
Algorithm Used:	Autoencoder and Tf-Idf measure
Input:	Gujarati Corpus
Output:	Disambiguated Sense

Steps to Gujarati Word Sense Disambiguation

1. **Preprocessing:** Collect large dataset of Gujarati Wordnet One by one read each paragraph of file.
2. **Tokenisation:** breaking line into array of tokens.
3. **Stop word removal**: remove high frequency words, you can use your own list of stop words.
4. Select best representative words as features using minimum frequency and maximum frequency threshold testing Use selected features to form term frequency vectors of each paragraph and save it.
5. Train deep autoencoder on a large dataset of different topics.
6. Convert the feature vector of a paragraph which (includes the ambiguous word) to the latent vector.
7. Now for each possible sense of ambiguous word, find mean latent vectors for each possible sense from Gujarati Wordnet.
8. Find cosine similarity between unknown latent vector and possible sense vector.
9. Sense having the highest cosine similarity is predicted as correct sense.

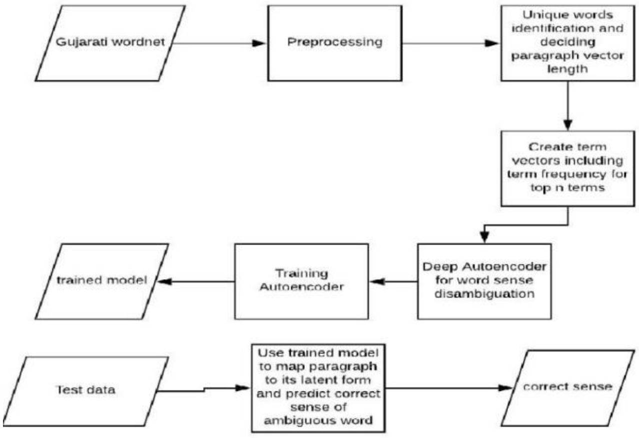

Gujarati wordnet
Preprocessing
Unique words identification and deciding paragraph vector length
Create term vectors including term frequency for top n terms
trained model
Training Autoencoder
Deep Autoencoder for word sense disambiguation
Test data
Use trained model to map paragraph to its latent form and predict correct sense of ambiguous word
correct sense

Figure 5 Proposed approach for Gujarati word sense disambiguation

VI. CONCLUSION

This model compares contextual similarity between an example paragraph and unknown paragraph. It contains a word having for which we want to find correct sense . This model depends upon the words selected as features in input vectors. For good generalization use more and more training sample and select best features covering most of the vocabulary. Use of colloquial terms in Gujarati language is much more than Hindi and it also varies from region to region. Gujarati alphabet is more difficult than English so it is possible to store same word with different spelling. It is also possible that due to inadequate knowledge of Gujarati language, some information can be stored with wrong spellings. Here solutions are proposed but they are just a starting step in the vast majority of the approaches are yet to be explored in Gujarati word Sense Disambiguation.

REFERENCES

1. Satanjeev Banerjee and Ted Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *International conference on intelligent text processing and computational linguistics*, pages 136{145. Springer, 2002.
2. CK Bhensdadia, Brijesh Bhatt, and Pushpak Bhattacharyya. Introduction to gujarati wordnet. In *Third National Workshop on IndoWordNet Proceedings*, volume 494, 2010.
3. Pranjal Protim Borah, Gitimoni Talukdar, and Arup Baruah. Approaches for word sense disambiguation {a survey. *International Journal of Recent Technology and Engineering*, 3(1):35{38, 2014
4. Niladri Sekhar Dash, Pushpak Bhattacharyya, and Jyoti D Pawar. *The WordNet in Indian Languages*. Springer, 2017.
5. Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24{26. Citeseer, 1986.
6. George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39{41, 1995.
7. George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, 3(4):235{244, 12 1990.
8. Andrea Moro, Francesco Cecconi, and Roberto Navigli. Multilingual word sense disambiguation and entity linking for everybody. In *International Semantic Web Conference (Posters & Demos)*, pages 25{28, 2014.
9. M Nameh, SM Fakhrahmad, and M Zolghadri Jahromi. A new approach to word sense disambiguation based on context similarity. In *Proceedings of the World Congress on Engineering*, volume 1, 2011.
10. SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wangchun Woo. Convolutional lstm network: A machine learning approach for precipitation now casting. In *Advances in neural information processing systems*, pages 802{810, 2015.
11. Torsten Zesch, Christof M¨uller, and Iryna Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *LREC*, volume 8, pages 1646{1652, 2008.

AUTHORS PROFILE

Prof Tarjni Vyas is working as an Assistant Professor in Computer Science and Engineering Department. She has more than 8 years of teaching experience. Prof Vyas received her BE degree in Computer Engineering from Gujarat University in 2009 and MTech degree in Computer Engineering from Dharamsingh Desai University (DDU), Nadiad in 2011.

Currently, she is pursuing her PhD in Computer Engineering from Charotar University of Science and Technology, Changa. Her area of specialization includes Natural Language Processing and Information Retrieval. She has good quality publications in international conferences and journals to her cedit. She is also involved in two research projects funded by SAC-ISRO under RESPOND and NISAR Scheme.

Dr. Amit P. Ganatra has received his B.E degree in Computer Engineering from Gujarat University, Gujarat, India in 2000 and master Degree from Dharmsinh Desai University, Gujarat, India in 2004. He has joined his Ph.D in the area of Multiple Classifier System (Information Fusion) at Kadi Sarvavishvidhalaya University, Gandhinagar, India in August 2008. Since 2000 he has been with faculty of Engineering and Technology, Charotar University of Science and Technology, Changa, Gujarat,where currently he is working as a Dean of Faculty Technology and Engineering, Devang Patel Institute of Advance Technology & Research (DEPSTAR), Charotar University of Science and Technology. He has published more than 150 research papers in the field of data mining and Artificial Intelligence. His current research interest includes Multiple Classifier System and Sequence Pattern Mining.