# Introduction to Gujarati wordnet

## Abstract

Gujarati is one of the 22 official languages of India. It is an Indo-Aryan language descended from Sanskrit. Gujarati wordnet is being built using expansion approach with Hindi as the source language. This paper describes experiences of building Gujarati wordnet. Paper discusses basic features of Gujarati language and evaluates suitability of Hindi language for expansion approach. Various issues related to synset linking using expansion approach and challenges related to language specific concepts are also discussed.

## 1 Introduction

Wordnets have emerged as a very useful resource for computational linguistics and many natural language processing applications. Since the development of Princeton WordNet (Fellbaum C., 1998), wordnets are being built in many other languages. 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. This paper describes experiences of building Gujarati wordnet.

The paper is organized as follows, section 2 gives introduction to Gujarati language, section 3 discusses the basic features of Gujarati language and section 4 describes influence of other languages on Gujarati and justifies use of Hindi language as base language for Gujarati Wordnet development. Synset development approach and synset categorization are discussed in Section 5 and 6 respectively. Section 7 gives the current status of Gujarati wordnet.Issues related to synset linking are discussed in section 8.

## 2 Gujarati Language

Gujarati, a native language of Indian state of Gujarat, is a member of Indo-Aryan family of languages. There are over 50 million speakers of Gujarati language.

Initially, the writing system of Gujarati was restricted to business writing , while the literature was in Devanāgarī script. The poetry form of the language is much older, enriched by poetry of poets like Narsinh Mehta. Gujarati prose writing and journalism started in 19th century. Protest writing against colonialism led to a string of powerful essays leading to the foundation of modern Gujarati literature.

## 3 Features

Some features of Gujarati language are as follows:

### 3.1 Writing system

Gujarati script is a variant of Devanāgarī script, differentiated by the loss of the characteristic horizontal line running above the letters and by a small number of modifications in the remaining characters.

For e.g.
Hindi: कमल, (kamal), Gujarati: કમળ

### 3.2 Vocabulary

As Gujarati is an Indo-Aryan language descended from Sanskrit, it's vocabulary contains four general categories of words: *tatsama, tadbhava*, *deshi* and *videshi* words.

- *tatsama*: Set of words accepted from Sanskrit language.
- *tadbhava*: Set of words from Sanskrit language adopted with a change in the phonological form.
- *deshi*: Words which are specific to Gujarati Language.
- *videshi*: Words which are accepted from different languages, like Persian, English, Portugese etc.

It is also noteworthy that in some cases *tatsama* and *tadbhava* words for a Sanskrit word co-exist with same or different meanings. For e.g. (1) ધર્મ ( Dharma) and ધરમ (Dharam) both means same, 'Religion'. While, (2) કર્મ (karma) means Work, with religious connotation and કરમ (karam) means Work in general sense.

### 3.3 Grammar

Gujarati follows Subject-Object-Verb word order. There are three genders and two numbers. There is no article. Some significant features are as follows:

#### 3.3.1 Gender

Gujarati distinguishes between three genders, masculine, feminine and neutral. For e.g.
છોકરો (chhokaro , Boy)
છોકરી (chhokarI , Girl)
છોકરૂ (chhokarU, Small kid)

However gender markers do not always represent the biological gender.
મંકોડો (mankodo , Big Ant)
મંકોડી (mankodI , Small Ant)

#### 3.3.2 Adjective

Adjectives agree with nouns and genders. A feminine adjective does not take plural marker while agreeing with a plural noun with feminine gender. For e.g.
Masculine singular
સારો છોકરો ('saro chhokaro' , Good Boy)
Masculine plural
સારા છોકરાઓ('sara chhokarao' , Good Boys)
Feminine singular
સારી છોકરી ('sari chhokari' , Good girl)
Feminine plural
સારી છોકરીઓ ('sari chhokario' , Good girls)

#### 3.3.3 Structure of verbs

Gujarati verbs have root+infinitive structure. Gujarati extends root verb to make causative sentence. For e.g.
ઝાડ પડ્યુ. ('Zaad padyu' , A tree fell)
રામે ઝાડ પાડ્યુ. ('raame Zaad paadyu' , raam caused the tree fell)
કાને રામ પાસે ઝાડ પડાવ્યુ. ('kaane raam paase Zaad padaavyu' , Kan cause Ram who caused the tree fell)

## 4 Influence of other languages on Gujarati

### 4.1 Comparison with Hindi

As an Indo-Aryan language, Gujarati language is very similar to Hindi. A brief comparison of Gujarati with Hindi is as follows,

- *Gender*: Gujarati language defines three genders while Hindi has only 2 genders.
- *Writing system*: Gujarati does not have the upper horizontal line running above the letter and few characters are modified.
- *Causative verbs*: Both Hindi and Gujarati handle causative verbs in the same fashion.
- *'Want' and 'should'*: Both Hindi and Gujarati handle "I should ..." and "I want .." in a similar ways. Gujarati uses 'jo' which is similar to 'chah' of Hindi.

  For e.g. ' I should go home now.' is written as,

  Hindi, 'मुजे घर जाना चाहीये।'

  Gujarati, 'મારે ઘરે જવુ જોઇએ.'

  (mare ghare javu joiAe)

### 4.2 Influence of other languages

There are other languages which also influence Gujarati. As India was ruled by Muslims, English and Portuguese, there is influence of these languages on Gujarati.

- *Urdu influence*: Following words demonstrate Urdu influence on Gujarati,

  દાવો (Urdu: dava English: Claim )

  ફાયદો (Urdu: fayda English: Benefit)

  કાયદો (Urdu: kayda English: Law )

  ખરાબ (Urdu: kharab English: Bad )

- *English influence*: Most of the Indian languages have adapted many of the English words and Gujarati is not an exception in that. For example,

  બેંક : Bank

  ફોન : Phone

  ટેબલ : Table

- *Portuguese influence*: Some of the words of Portuguese language adapted in Gujarati are as follows,

  સાબુ : 'saabu' soap

  બટાટા : 'bataataa' potato

  પાદરી : 'paadarI' father (Christian priest)

Thus, Gujarati language has rich set of words derived from Indian languages as well as foreign languages. This insight helps in selecting an approach for building wordnet.

## 5 Synset Development Approach

Gujarati wordnet is being built using expansion approach (Vossen P., 1998). In this approach synsets are created by referring to existing wordnet of related language. Hindi is used as a source language to create synsets of Gujarati language. Benefits of this approach are: (1) Wordnet development process becomes faster as the gloss and synset of the source language is already available as reference. (2) It provides linking between the synsets of different languages which can be used for machine translation applications.

The task of synset development for Gujarati language is further simplified by availability of the on line lexical resources like *'Bhagavad Go Mandal'* (Patel C. B.(ed) , 1958) and 'Gujarati Lexicon' (Chandaria R. , 2006). 'Bhagavad Go Mandal' contains around 8.2 lacs words spread across 9 volumes. 'Gujarati Lexicon' is an another more recent effort. The online interface of Gujarati lexicon provides easy access to meanings, synonyms, antonyms, idioms, proverbs and phrases. These two resources provide great help in building synsets.

As Gujarati language is closely related to Hindi, the most of Gujarati synsets are created by translating Hindi synsets to Gujarati synsets. However, emphasis was given to understand the concept independently of a language and then to create synset. Though notion of concept is defined independently of the language, many times it was observed that the concept present in Hindi was not present in Gujarati or even though the concept was present there was no indigenous lexeme for the concept.

## 6 Synset Categorization

As described in previous section, sometime, there is disagreement on concepts across languages. Many concepts of Hindi are not present in other languages or there is no indigenous lexeme for the concept in other language. So, to facilitate synset linking across languages, Hindi synsets are divided into following different categories,

- *Universal* : This set of concepts is present in all the languages and is essential and most frequently used. For e.g., 'सूर्‌य' (sun). Most of these concepts belong to top-level of the wordnet and are directly linked with English WordNnet and SUMO.

- *Pan-Indian* : This set of concepts is common in all Indian languages and linkable across all Indian languages but does not have parallel concept in English. For example, 'तबला' (tabala)(An Indian rhythm instrument).

- *In-Family* : These are the concepts common in specific subsets of Indian languages and linkable across all languages of the family. For example: 'चाचा' (chacha)(paternal uncle) 'भतिजा' (bhatija) (brother's son)

- *Language Specific* : These concepts are specific to a language. These concepts are specific to the culture. It includes local food, festivals,etc. For example, 'बीहु', (Bihu) (Name of festival celebrated in Assam state of India) word is very specific to the state and the culture and does not appear in any other language. These concepts appears very low in the hierarchy of the wordnet and normally represents instances or individuals.
- *Rare* : This includes very specific words adopted in most of the languages. It includes specific technical or scientific terms like, 'ngram'.
- *Synthesized* : These are the synsets created in a language due to the influence of other languages. These synsets are not natural to the language but needed to link synsets of two different languages.

Such classification of synsets helps in linking concepts of different languages. For example, if a synset belongs to the universal synset then it is present in both Hindi and English language. And if a synset belongs to the Pan-Indian category then it belongs to both Hindi and Gujarati languages. Thus, wordnet development using expansion approach will be faster by this method.

Till date, 7163 universal synsets and 1356 Pan-Indian synsets have been manually identified and are now linked across all languages. Out of 7163 universal synsets 7012 are directly linked with English wordnet synset and 24 are linked through hypernymy. Out of 1347 Pan-Indian synsets 287 are directly linked with English wordnet synset and 125 are linked through hypernymy. The 24 Universal synsets represent the concepts which are not present in a specific Indian language. 287 directly linked synsets represent concepts which are adopted in English language. Language specific synsets are being developed and then they will be linked by translating them into Hindi and English.

## 7 Synset Development status

Till date, 15595 synsets, covering 42537 words, are built in the Gujarati wordnet. The category-wise count of synsets is as follows:

| | |
|---|---|
| Universal: | 7169 |
| Pan-Indian: | 1348 |
| Language specific: | 108 |
| Verb: | 1799 |
| Adverb: | 210 |
| Adjective: | 3606 |

## 8 Issues related to synset development

During the development of synsets, some disagreements were observed between Hindi concepts and Gujarati concepts.

### 8.1 Hindi synsets not linked with Gujarati

Following are some examples of Hindi synsets not linked with Gujarati,

- *Difference in concept description*

  Concept: तुरही की तरह का एक बड़ा बाजा

  Example: "नरसिहा की आवाज़ दूर–दूर तक सुनाई देती है"

  Synset: नरसिहा, नरसिगा, गोमुख

  No such concept is identified in Gujarati language. However, there is a concept in Gujarati language for similar instrument which is used at war-front to announce beginning of a war.

- *No indigenous lexeme in Gujarati*

  Concept: इत्र का व्यापार करनेवाला व्यक्ती

  Example: "आजकल, इत्र व्यापारी नक़ली इत्र का व्यापार भी करने लगे हैं"

  Synset: इत्र व्यापारी, इत्र फरोश, इत्र फ़रोश, अत्तार, गंधी, गन्धी

  There is no indigenous lexeme for this concept in Gujarati language.

- *Confusing gloss*

  Concept: एक छोटा पक्षी जो प्रायः अपना घोसला मकानो में बनाता है

  Example:"गौरैया अपने बच्चो को दाना चुगा रही ह"

  Synset: गौरैया, गौरेया, वृषायण, आकली

  The concept is general and exists in Gujarati language but it is difficult to identify the Gujarati name of the bird from the synset.

- *Difficult to adopt*

  Concept: जो प्रवीष्ट न हुआ हो

  Example: “अप्रवीष्ट महेमानो को शीघ्र ही भीतर प्रवेश करने दे”

  Synset: अप्रवीष्ट

  Though this word can be translated in Gujarati language, it is not a native concept used in Gujarati language.

- *No such concept in Gujarati*

  Concept: जो अकेला चरता या वीचरण करता हो

  Example: “जंगली सूअर एक पृथकचर पशु है”

  Synset: पृथकचर

  There is no such concept in Gujarati language.

Concept described above are not part of general vocabulary and represent very specific nouns. There was no difficulty in linking verb, adjectives or causative verbs. This is due to the similarity between Hindi and Gujarati languages. Out of around 7800 concepts of Hindi language referred so far, around 7500 concepts were linked to Gujarati language.

### 8.2 Language specific synset

While major part of the day to day vocabulary of Gujarati language is similar to that of Hindi, there are some concepts which are very specific to Gujarati language. These concepts are very specific to the culture of Gujarat. These concepts refer to food items, places, traditions, religion etc. Some of the examples are as follows:

- *Culture specific concept*

  Concept : કોઈ ખાસ પ્રસંગે કસુંબો પીવા માટે ભેગા થવું

  Example : “ગુજરાત ના કોઈ ગામો માં આજે પણ ડાયરા થાય છે”

  Synset : ડાયરો (डायरो, Daayaro)

- *Tradition specific concept*

  Concept : એક ફળ કે જે લગ્ન પ્રસંગે વર કન્યા ના હાથે બાંધે છે

  Example : “લગ્ન પછી વર કન્યા મીંઢળ છોડે છે”

  Synset : મીંઢળ (मींढळ, mIMdhaL)

- *religion specific concept*

  Concept : મોક્ષ માટે ભગવાન નું નામ લેતા લેતા ગીરનાર પર થી પડતું મુકવું.

  Example : “ગીરનાર શીખર પર થી ભક્તો ભૈરવજપ કરતા હતા”

  Synset : ભૈરવજપ (भैरवजप, bheiravajapa)

## 9 Conclusion

Existence of Hindi wordnet and similarity between Hindi and Gujarati languages helped development of Gujarati wordnet. Also, the resources like 'Bhagavad-Go-Mandal' and 'Gujarati Lexicon' were found to be very useful in synset development process. Synset categorization further simplified the synset linking process. It is observed that most of the top level concepts are common and easily linked. The concepts that vary across languages are specific to culture and tradition of the people. Mostly these are noun concepts and do not have hyponymy. Many of these are singleton synsets that appear very low in the wordnet concept hierarchy. The future work is to identify and link language specific and in-family concepts. It is also required to develop lexical relations and to evaluate suitability of semantic relations of Hindi wordnet for Gujarati language.

## References

Christiane Fellbaum 1998. *WordNet: An Electronic Lexical Database.* MIT Press

Piek Vossen 1998. *EuroWordNet: a multilingual database with lexical semantic networks.* Kluwer Academic Publishers

D. Chakrabarty P. Pande D. Narayan and P. Bhattacharyya 2002. *An experience in building the Indo WordNet - a WordNet for Hindi* International Conference on Global WordNet (GWC02), Mysore, India

Patel C. B. 1958. *Bhagvad-Go-Mandal.* http://www.bhagavadgomandalonline.com.

Ratilal Chandaria 2006. *Gujarati Lexicon* http://www.gujaratilexicon.com