Introduction to Gujarati Wordnet

Prof. C. K. Bhensdadia
ckbhensdadia@ddu.ac.in
Department of Computer Engg.,
Dharmsinh Desai University, Nadiad

Brijesh Bhatt
brijesh@cse.iitb.ac.in
Department of Computer Science and Engineering,
Indian Institute of Technology, Mumbai

Prof. Pushpak Bhattacharyya
pb@cse.iitb.ac.in
Department of Computer Science and Engineering,
Indian Institute of Technology, Mumbai

Abstract
Gujarati language is the youngest member of IndoWordnet[1]. As a part of IndoWordnet project, Wordnet for Gujarati language is being developed from Hindi Wordnet using expansion approach. This paper reviews the Gujarati Wordnet development process. It describes the basic features of Gujarati language and evaluates suitability of Hindi language as a source language. Also, the current status of the work and the issues in development are described.

1. Introduction

WordNet[2] is a machine readable lexical database for English language developed at Princeton University. It has evolved as the most valuable resource for the natural language processing application. Following the Princeton WordNet, wordnets for many other languages were developed across the globe. The first wordnet for Indian languages is Hindi wordnet[3], developed at Indian Institute of Technology, Bombay. Recently, efforts are going on to develop wordnets for many Indian Languages. One such effort is to build Gujarati wordnet from Hindi wordnet using expansion approach.

The layout of the paper is as follows: section 2 gives introduction to Gujarati language, section 3 describes historic influence of other languages on Gujarati and justifies use of Hindi language as base language for Gujarati Wordnet development. Section 4 describes the expansion approach selected for the Wordnet development. Section 5 describes the status of Gujarati Wordnet devleopment and some issues related to synset linking.

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 and it is one of the 22 official languages of India. Incidentally, Gujarati was the first language of Gandhiji (Mohandas K. Gandhi, father of India) and Mohammed Ali Jinnah (father of Pakistan).

2.1 History

Initially, the writing system of Gujarati was restricted to business writing , while the literature was in Devanāgarī script. The poetry form of 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.

2.2 Features

Some features of Gujarati language are as follows:

2.2.1 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 example:

Hindi: कमल
(kamal)
Gujarati: કમળ

2.2.2 Vocabulary: As Gujarati is an Indo-Aryan language descended from Sanskrit, it's vocabulary contains four general categories of words:
Tatsam, Tadbhav and Native and Loan words.

***Tatsam*:** Set of words accepted from Sanskrit language.

Tadbhav: Set of words from Sanskrit language adopted with change in phonological form.

Native: Words which are specific to Gujarati Language.

Loan Words: Words which are accepted from different languages, like Persian, English,

Portugese etc. Next section describes such words in more detail.

It is also noteworthy that in some cases *tatsam* and *tadbhav* words for same Sanskrit word co-exist with same or different meanings.

For example:

(1) ધર્મ (Dharma) and ધરમ (Dharam) both means same, 'Religion'.

(2) કર્મ (karma) : Work, with religious connotation
કરમ (karam) : Work

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

2.2.3.1 Gender: Gujarati distinguishes between three genders : masculine, feminine and neutral. However the gender marker do not represent the biological gender all the time.

For example:

છોકરો	**છોકરી**
(chhokaro)	(chhokari)
(Boy)	(Girl)

મંકોડો	**મંકોડી**
(mankodo)	(mankodi)
(Big Ant)	(Small Ant)

2.2.3.2 Adjective: Adjective agrees with noun and gender. Feminine adjective does not take plural marker while agreeing with a plural noun with feminine gender.

For example:

(1) Masculine singular
સારો છોકરો
(sar-o chhokar-o)
Good Boy

(2) Masculine plural
સારા છોકરાઓ
(sar-a chhokara-o)
Good Boys

(3) Feminine singular
સારી છોકરી
(sar-i chhokar-i)
Good girl

(4) Feminine plural
સારી છોકરીઓ
(sar-i chhokari-o)
Good girls

2.2.3.3 Structure of verbs: Gujarati verbs have root+infinitive structure. Gujarati extends root verb to make causative sentence.

For example:

(1) **ઝાડ પડ્યુ.**
(Zaad paDyu)
A tree fell.

(2) **રામે ઝાડ પાડ્યુ.**
(Rame Zaad paaDyu)
Ram caused the tree fell.

(3) **કાને રામ પાસે ઝાડ પડાવ્યુ.**
(Kane Ram paase Zaad padaVyu)
Kan cause Ram who caused the tree fell.

3. Influence of other languages on Gujarati

As an Indo-Aryan language, Gujarati language is very similar to Hindi, Marathi and Punjabi. Grammar and vocabulary of Gujarati language is very similar to Hindi with few exceptions. A brief comparison is as follows :

(1) Gender: As described in section 2, Gujarati language defines three genders while Hindi has only 2 genders.

(2) Writing system: Gujarati dropped the upper horizontal line running above the letter, and few characters are modified as shown in the previous section.

(3) Causative verbs: Both Hindi and Gujarati handle causative verbs in the same fashion.

For Example,

Hindi: रोना रुलाना रुलवाना
(rona) (rulana) (rulavana)

is similar to,

Gujarati: **રડવું રડાવવું રડાવરાવવું**
(radvu) (radavavu) (radavravavu)

(4) 'Want' and 'should': Both Hindi and Gujarati handles "I should ..." and "I want .." in similar ways. Gujarati uses 'jo' which is similar to 'chah' of Hindi.

For example,

I should go home now.

in Hindi,

मुजे घर जाना चाहीये।

in Gujarati,

મારે ઘરે જવુ જોઇએ.
(mare ghare javu joiAe)

However 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,

Gujarati	Urdu	English
દાવો	dava	Clami
ફાયદો	fayda	Benefit
કાયદો	kayda	Law
ખરાબ	kharab	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: Following are the some of the words of Portuguese language adapted in Gujarati:

સાબુ	soap
બટાટા	potato
પાદરી	father (Christian priest)

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

4. Gujarati Wordnet development using expansion approach

Gujarati wordnet is being built using expansion approach[4]. In this approach, instead of creating the synset from the scratch, synsets are created by referring to existing wordnet of related language. Hindi is used as a source language to create synsets of Gujarati language. The 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.

Synset linkage tool, provided by I.I.T.Bombay, is used to create synset of Gujarati language. This synset linking tool provides graphical user interface which shows Hindi synset on the left side and provides interface to enter Gujarati synset on the right hand side.
As Gujarati language is closely related to Hindi, most of the Gujarati synsets are created by translating the Hindi synset to Gujarati synset. However, emphasis was given to understand the concept independently of language and then to create synset.
The task of synset development for Gujarati language is further simplified by on line availability of the milestone laxicon resources like *'Bhagavad Go Mandal'*[5] and *'Gujarati Lexicon'*[6]. '*Bhagavad Go Mandal*' was created in early twentieth century at princely state of Gondal in Kathiawad. It contains around 8.2 lacs words spread across 9 volumes. It is accepted as standard reference for Gujarati language by '*Gujarat Sahitya Parishad*' under the leadership of Mahatma Gandhi. '*Gujarati Lexicon*' is an another more recent effort, by Ratilal Chandaria. 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.

5. Observations

5.1 Synset linkage status

The synsets are divided into two categories- Core and Common. Following is the status of synset developed under each categories.

Core synset
No. of synsets: 1866
Total words : 7985
Unique words: 7078

Common synset
No. of synset : 5632
Total words : 17245
Unique words: 13800

5.2 Issues related to synset development

Some Hindi synsets were not linked with Gujarati synsets because of the following reasons:
(1) Concept does not exist in Gujarati language
(2) Difficulty in interpreting gloss of Hindi synset.
Some examples are as follows:

Core synset

(1) ID: 408

Concept: तुरही की तरह का एक बड़ा बाजा

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

Synset: नरसिंहा, नरसिंगा, बाँकिया, गोमुख, सिंगा

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.

(2)ID: 2636

Concept: इत्र का व्यापार करनेवाला व्यक्ति

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

synset: इत्र व्यापारी, इत्र फरोश, इत्र फ़रोश, अत्तार, गंधी, गन्धी, इत्रफ़रोश, इत्रफरोश, इत्रफ़िरोश, इत्रफिरोश

There is no such concept in Gujarati language.

(3) ID: 4436

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

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

Synset: गौरैया, गौरेया, स्वल्पघटक, वृषायण, बहुशुत्र, 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.

Common synset

(4) ID : 3

Concept: जो प्रविष्ट न हुआ हो

Example: "अप्रविष्ट अतिथियो को शीघ्र ही भीतर प्रवेश करने दिया जाय"

Synset:अप्रविष्ट

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

(5) ID : 613

Concept: जो अकेला चरता या विचरण करता हो

Example: "जंगली सूअर एक पृथकचर पशु है"

Synset: पृथकचर

There is no such concept in Gujarati language.

So, above examples describes some of the synsets for which Gujarati synsets couldn't be created. However, these synsets are not part general vocabulary.

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 which means over 95% concepts are common to both languages.

5.3 Gujarati language specific concepts

While most of the 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 mostly related to unique features of Gujarati language and Gujarati literature. Some of the examples are as follows:

(1) **ગરબો** (Garabo): Sacred light to worship Goddess during *Navratri.*

A form of dance performed by women to worship goddess during *Navratri.*

(2) **ભવાઇ** (BhavaI): Specific form of drama, with special characters like *'Ranglo'* and *'Rangli'* used in ancient days to convey the social issues. Though a rare form of an art, the concept is still very common to Gujarati language.

(3) **છપ્પા** (Chhappa): A specific from of poetry, similar to *'Dohe'*. However, it is different from *'Doha'* as it exists separately in Gujarati language.

6. Conclusion

Existence of Hindi wordnet and similarity between the Hindi and Gujarati language 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. Effort of developing wordnet using expansion approach for various Indian language is going to produce huge lexicon resource which will prove to be invaluable for machine translation and Natural Language processing applications.

Acknowledgement

This work is done under project 'Indradhanush-Wordnet development project for seven Indian languages', sponsored by Ministry of Communication & I. T., India. We sincerely acknowledge DIT for providing support for the project.

References

[1] Bhattacharyya P. (2009) "IndoWordnet", Lexical Resources Engineering Conference 2010 (LREC 2010), Malta, May, 2010.

[2] Fellbaum C. (1998) “WordNet: An Electronic Lexical Database.” MIT Press.

[3] Narayan D., (2002) "An Experience in Building the Indo WordNet- a WordNet for Hindi, 1st International Conference on Global WordNet (GWC 02), Mysore, India.

[4] Vossen P. (ed.). 1998 "EuroWordNet: A Multilingual Database with Lexical Semantic Networks." Kluwer Academic Publishers, Dordrecht.

[5]'Bhagvad-Go-Mandal', http://www.bhagvadgomandalonline.com

[6] 'Gujarati Lexicon', http://www.gujaratilexicon.com