Telugu WordNet

S. Arulmozi
Department of Dravidian & Computational Linguistics
Dravidian University, Kuppam 517425, India
arulmozi@gmail.com

Abstract

This paper describes an attempt to develop Telugu WordNet, particularly construction of synsets in Telugu language along the lines of Hindi synsets using the expansion approach. Based on the Hindi WordNet synsets, we assign Telugu synsets manually using the Offline Tool Interface. We share the challenges faced in the construction of core synsets from Hindi into Telugu language. A brief account on Telugu language and its notable features are also provided.

1 Introduction

WordNet building activities in Dravidian languages started with the work of Tamil WordNet[1] at AU-KBC Research Centre using Rajendran's (2001) ontological classification of Tamil vocabulary. Work on Dravidian WordNet (comprising WordNets in four major Dravidian languages, viz. Kannada, Malayalam, Tamil and Telugu) started during a Workshop[2] held at Chennai in which synsets were built for Construction Domain. Currently Dravidian WordNet[3] activity is being carried out for Kannada at University of Hyderabad, Malayalam at Amrita Vishwa Vidyapeetham, Tamil at Tamil University and Telugu at Dravidian University.

In this paper, we describe the construction of synsets for Telugu language. Based on the Hindi WordNet synsets, we aim to develop synsets in Telugu using the expand model. The paper is organized as follows: Section 2 gives a brief account on the morphological features of Telugu language. This section also provides the language technology activities pertaining to Telugu language. Section 3 details the Telugu synset building activity, challenges/problems faced during the construction of synsets. Section 4 gives a statistical account on the synsets developed. The last section summarizes the work.

2 The Telugu Language

Telugu belongs to the South Central Dravidian subgroup of the Dravidian family of languages. It has recorded history from 6th Century A.D. and literary history dating back to 11th Century A.D. It has been recently awarded the Classical Status. It is the second most spoken language after Hindi in India. Telugu has been the language of choice for lyrical compositions for its vowel ending words, rightly called the "Italian of the East".

The vocabulary of Telugu is highly Sanskritized in addition to the Persian-Arabic borrowings కబురు/ kaburu/ `story ', జవాబు /javaabu/ `answer '; Urdu తరాజు /taraaju/ `balance'. It does have cognates in other Dravidan languages such as పులి /puli/ `tiger ', ఊరు /uuru/ `village'; తల /tala/ `head'.

Words in Dravidian languages, especially in Telugu are long and complex. i.e. because it also suffixation, words are build up from many affixes that combine with one another. Telugu, like other Dravidian languages is highly rich in morphology and hence agglutinative in nature; it also allows polyagglutination.

Telugu has the state-of-the-art Morphological Analyser [4] and Telugu-Hindi, Hindi-Telugu, Telugu-Tamil, Tamil-Telugu MT systems[5]. In addition to this, the Resource Centre for Indian Language Technology Solutions-Telugu (RCILTS) established by the Ministry of Communications and IT, Govt.of India during 2000-2003 has developed several products,

[1] Project partially funded by Tamil Virtual University.
[2] Workshop on WordNet for Dravidian Languages organized from 2-3 June 2003
[3] Project funded by the Ministry of HRD, Govt. of India.
[4] Developed by G.Uma Maheswar Rao at the Centre for ALTS, University of Hyderabad.
[5] Consortia on Indian-Language-Indian Language MT systems established by MCIT, GoI.

services and knowledge bases pertaining to Telugu language. They include Drishti, the first comprehensive OCR system in Telugu, Tel-Spell, Spell checker for Telugu, Telugu Corpus (9.2 m words)[6], etc.

This effort on building a WordNet for Telugu is the first of its kind along the lines of Hindi WordNet.

3 Construction of Telugu WordNet

Various approaches are followed in the construction of WordNets across the languages of the world. For the Indian languages, WordNets are constructed using the expand model.

For the construction of Telugu WordNet, we also follow the expand model. i.e. Hindi WordNet synsets are taken as a starting point of departure. The concepts provided along with the Hindi synsets are first conceived and appropriate concepts in Telugu are manually provided by language experts. The Telugu synsets are then built based on the concepts created keeping in view the three principles, viz. Minimality, Coverage and Replaceability.

For the building of synsets, we use the OfflineTool provided along with Hindi WordNet synsets. This standalone interface allows users to view the Hindi synsets, concepts, example sentence on the one side and simultaneously keying the target language (Telugu in our case) synsets, concepts and example sentence. The tool also has the Princeton WordNet English synsets interlinked. This helps the language experts to cross check with English WordNet synsets.

Below we present the challenges/problems faced in the construction of 2000 core synsets from Hindi into Telugu.

3.1 Expansion from Hindi/English into Telugu

For the construction of Telugu synsets, we use the OfflineTool provided by Hindi WordNet group. We have encountered certain problems with regard to the tool. We have also faced specific problems while rendering corresponding synsets, concepts in Telugu. Below we present few of the computational and lexicographical concerns.

3.1.1 Computational Concerns

The interface[7] has the following problems:

1. In the target language box, complete option does not show how many synsets are completed. It was working in the earlier version.

2. By mistake, if the user opens the tool for the second time, there shall be a message displayed stating, that the tool is already open.

3. The tool does not work when logged-in as a Guest user (in Windows Vista). Whenever guest user is logged in and opens the tool, it does not allow to SAVE the data.

4. When X-Link (Synset Linker) is invoked, the popup window does not display any options.

5. Everytime the user has to use the exit button to close the tool, instead the close option X shall be enabled.

3.1.2 Lexicographical Concerns

1. Hindi synsets ID 5499:

Synset IDs are the same in both versions (2.0 and 2.1) of Offline Tool. Subsequently, the concept, example and sysnets are the same.

But in version 2.0 of the OfflineTool, EWN ID is 1214525 and it displays the Concept as: “the social act of assembling”; Example as: "they demanded the right of assembly"; Synset: “assembly, assemblage, gathering” and in version 2.1, PWD ID is 8069519 displaying the Concept as: “a large number of things or people considered together”; Example: "a crowd of insects assembled around the flowers"; Synset: “crowd”.

Now, this is a problem when one tries to manually assign synsets in Telugu based on the Hindi/English synsets.

2. Hindi ID Number 7379:

The concept भाई का लड़का /bhai ka ladka/ which means `brother’s son’ is an interesting problem. The synsets given for this concept in Hindi are

[6] Achievements of RCILTS-Telugu, TDIL

[7] Offline Tool Version 2.1

“भतीजा, भ्रातृज, भ्रातापुत्र, भ्रातृपुत्र, भतीज, अवतंस, अवतन्स” /BatIjA, BrAtRuja, BrAtAputra, BrAtRuputra, BatIja, avataMsa, avatansa/ which all means `brother’s son’ in Hindi. When it comes to Telugu, providing concept is a challenge. Straightforward, one can assign సోదరుని కుమారుడు /sOdaruni kumAruDu/. But when assigning synsets, one has to cope up with the ambiguity in the concept. i.e. whose brother;s son.
In any case, we have provided equivalent as అన్న కొడుకు /anna koDuku/.

Similar challenge faced in ID 1804.

3. Hindi synset ID: 7531

Concept:
चालीस सेर की एक तौल
cAlIsa sera kI eka taula

which means `a measure of 40 kg’.

For this concept, there is no corresponding equivalent in Telugu. But there are different usages in different dialects of Andhra Pradesh. For example, in Kuppam, the measure is equal to 10 kg whereas in Kadapa it is 14 kg.

When it comes to providing equivalent synsets, Hindi and Telugu uses the same, i.e. *manu*.

4. Hindi Synset ID: 1602

Concept:
व्यापार करनेवाला व्यक्ति
vyApAra karanevAlA vyakti

which means `bookbinder’

The English IDs given for the above mentioned Hindi synset ID is different from Version 2.0 to Version 2.1.

In Version 2.0 of OfflineTool, EWN ID is 10560080 and the concept given is “someone who purchases and maintains an inventory of goods to be sold”, and synsets are “trader, bargainer, dealer, monger”.

In Version 2.1, under PWN ID 9736400, the concept is “a person engaged in commercial or industrial business (especially an owner or executive) and the synsets are “businessman, man of affairs”.

5. Hindi Synset ID: 1680

Concept:
बहुत बड़ा या विशेष ऊँचाई का या जिसका विस्तार ऊपर की ओर अधिक हो
bahuta baDxA yA viSeSha UMcAyI kA yA jisakA vistAra Upara kI ora adhika ho

For the above mentioned Hindi concept, the corresponding English synsets in both the versions are completely different.

In the 2.0 version, for the EWN ID 1250892, the concept given is “the action of establishing on a socialist basis” ; example given is “the socialization of medical services” and the synsets are “socialization, socialisation”.

In 2.1 version, for the PWN ID 1250892, the concept given is “(literal meaning) being at or having a relatively great or specific elevation or upward extension (sometimes used in combinations like “knee-high””, example is “a high mountain”, and so on and the synset is “high”.

These kinds of synsets pose a real problem for the language experts because whenever they assign a concept/synset, it creates a confusion. Further, it

We have presented only few challenges and problems faced in the construction of Telugu WordNet using Hindi synsets. The real challenge will come up once we start working horizontally on synsets.

4 Statistics

The status of Telugu WordNet after completion of core (2k) synsets is given below:

No. of Core Synsets: 2000
No. of unique words: 4270

Out of 2000 core synsets in Telugu, 3489 are noun synsets with synsets ranging from 1-11; 521 are verb synsets ranging from 1-6; 498 adjective synsets ranging from 1-8 and 145 adverb synsets ranging

from 1-7. The unique words from all the POS categories for the 2000 core synsets is 4270.

	Nouns	**Verbs**	**Adj**	**Adv**
Synsets	3489	521	498	145
1	495	50	34	8
2	454	90	53	15
3	334	58	43	16
4	131	21	31	9
5	53	3	12	2
6	13	3	5	1
7	15	0	1	1
8	8	0	1	0
9	3	0	0	0
10	1	0	0	0
11	1	0	0	0

Table 1: Status of Telugu Synsets

5 Summary

In this paper, we have described the approach that is followed for the construction of Telugu WordNet. A brief note on the characteristics of Telugu language and the language technology activity in Telugu is also provided. The manual synset building activity of 2000 core synsets with specific problems faced is discussed. Work is in progress for completing 10k common synsets and this will be made available during the presentation.

Acknowledgements

The work on Telugu WordNet activity is part of the larger effort on building a Dravidian WordNet and funded by the Ministry of Human Resource Development, Government of India.

References

Brown, C.P. 1857. *A Grammar of the Telugu Language*. Christian Knowledge Society's Press, Madras.

Burrow, T. and Emeneau, M. B. 1984. *Dravidian Etymological Dictionary*. Munshiram Manoharlal Publishers, New Delhi.

Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Krishnamurti, Bh. 1961. *Telugu Verbal Bases*. University of Chicago Press, Berkeley.

Krishnamurti, Bh. And Gwynn, J.P.L. 1985. *A Grammar of Modern Telugu*. Oxford Univeristy Press, New Delhi.

Krishnamurti, Bh. 2003. *The Dravidian Languages*. Cambridge University Press, Cambridge.

Narayan, D., Chakrabarty D., Pandey P. and Bhattacharyya, P. 2002. `An Experience in Building the Indo WordNet- a WordNet for Hindi', *International Conference on Global WordNet*, Mysore.

Miller, G.A. 1995.` WordNet: A Lexical Database for English', *Communications of the ACM*. Vol.38, No.11.

Rajendran, S. 2001. *taRkaalat tamizhc coRkaLanjciyam* [Modern Tamil Thesaurus]. Tamil University Publication, Thanjavur.