Experiences in building the Urdu WordNet

Farah Adeeba
Center for Language Engineering
Al-Khawazmi Institute of Computer Science
University of Engineering and Technology
Lahore
farah.adeeba@kics.edu.pk

Sarmad Hussain
Center for Language Engineering
Al-Khawazmi Institute of Computer Science
University of Engineering and Technology
Lahore
Sarmad.hussain@kics.edu.pk

Abstract

This paper attempts to report on developing a WordNet for Urdu on the basis of Hindi WordNet. The resource currently contains about 50000 unique words organized in 28967 synsets. The paper also discusses the problems encountered along the way of transliteration from Hindi WordNet and manual cleaning. It concludes with the planned future work.

1 Introduction

WordNet is one of the useful and important lexical resources based on the formalisms developed in lexical semantics. It defines different senses associated with the meaning of a word and other well-defined lexical relations such as synonyms, antonyms, hypernym, hyponyms, meronyms and holonyms. WordNet is used for many natural language processing and computational linguistic tasks such as Word Sense Disambiguation, Word Similarity, Information Retrieval and Extraction and Machine Translation, etc.

The motivation for the creation of Urdu WordNet is to provide a lexical resource that can be used as a tool for enhancing the performance of machine translation and information retrieval. We have attempted to provide a basic resource that can be used in above mentioned NLP applications. As the manual construction of Urdu WordNet from scratch would be very costly and time consuming, we have used the WordNet expansion approach. Lexical information is extracted from Hindi WordNet due to similarity between two languages.

Hindi and Urdu are grammatically similar languages but written in two dissimilar scripts Devanagri and Arabic respectively. These languages share a large number of words, morphology, vocabulary, and cultural heritage. It is easier for both speakers to verbally understand each other but they face the barrier of different script incase of written expression. Hindi and Urdu are spoken by more than 60 million people in India and Pakistan (Language Summary, http://www.ethnologue.com/ethno_docs/distribution.asp?by=size).

The roadmap for the rest of paper is as follows: Section 2 discusses Hindi and Urdu Scripts along Hindi WordNet. Methodology for development of Urdu WordNet is described in Section 3 and statistics of system is given in Section 4. The current status and future work is discussed in Section 5. Finally section 6 concludes the paper.

2 Literature Overview

Urdu (اردو) is written in Persio-Arabic script and normally in Nastaliqb writing style (Hussain, 2004). It is a right-to-left script and the shape of character differ depending on its position in word i.e. shape of character would be different in initial, middle, and end of word. Urdu is written in bidirectional form i.e. letters are written from right-to-left and numbers from left-to-right format. Urdu is written with consonantal letters and aerabs. The vocalic content is specified by using the aerab with letters. Aerab position can be on the top and bottom of letter. A sentence illustrating Urdu is given below:

اردو عربی رسم الخط میں لکھی جاتی ہے۔

(Urdu Arbi Rasm-ul-Khat mein likhi jati hay)
(Urdu is written in Arabic script)

Hindi (हिन्दी) is written in Devanagri script, descended from the Brahmi script. It is the simplified version of Sanskrit, written in left-to-

Proceedings of the 9th Workshop on Asian Language Resources, pages 31–35,
Chiang Mai, Thailand, November 12 and 13, 2011.

right direction. In Hindi each consonant letter by default inherits vowel which can be altered or muted by means of diacritics or matra. Vowels can be written as independent letters or by using a diacritic marks. Two or more consonants may occur together in clusters called Conjunct. A sentence written in Hindi is given below.

हिन्दी हिंदूस्तान की कौमी ज़बान है.
(Hindi India ki Quomi Zuban hay)
(Hindi is the national language of India)

Hindi WordNet (HWN) is lexical database inspired by the English WordNet (Miller, 1993).The words in HWN are grouped together according to their similarity of meanings. Two words that can be interchanged in a context are synonymous in that context. Synsets or the synonym sets are the basic building blocks of HWN. For each word there is a synonym set, or synsets representing one lexical concept. There are 10 relations in HWN; Synonymy, Hypernymy / Hyponymy, Antonymy, Meronymy / Holonymy, Gradation, Entailment, Troponymy and Causative (Dipak , 2002). The Hindi WordNet deals only with the open class words. Thus, HWN contains the following categories of words. The details of Hindi WordNet are given in Table 1.

Category	Count
Noun	56623
Verbs	3894
Adjectives	13702
Adverbs	1276
Synsets	30977

Table 1: Hindi WordNet

Hindi WordNet is used as a pivot WordNet for building WordNet of Ando-Aryan languages eg. Marathi WordNet, Sanskrit WordNet (Kulkarni, 2010), Nepali WordNet (Chakrabarty, 2006) , Bengali WordNet. The Expansion Approach of WordNet (Vossen, 2002) creation is used as method for creation of a new Word Nets. This expansion approach is also being used for development of Urdu WordNet by Tafseer et.al (Tafseer Ahmed, 2010).

They developed Urdu WordNet by extracting information contained in existing Hindi WordNet. To overcome the scriptural barrier they used transliteration. The lexical information is obtained by using the Hindi WordNet API. The gloss with the example sentence and the synset description is left out into Urdu WordNet.

3 Methodology

Because of the high degree of similarity between the Urdu and Hindi, we have picked up the Hindi WordNet as the pivot WordNet. The HWN offline version of 2.1 is being used that provides information of synset and senses. The Hindi WordNet database is picked up from (http://www.cfilt.iitb.ac.in/wordnet/webhwn/downloaderInfo.php) and transliterated it into Urdu. Hindi to Urdu WordNet conversion process is shown in Figure 1.

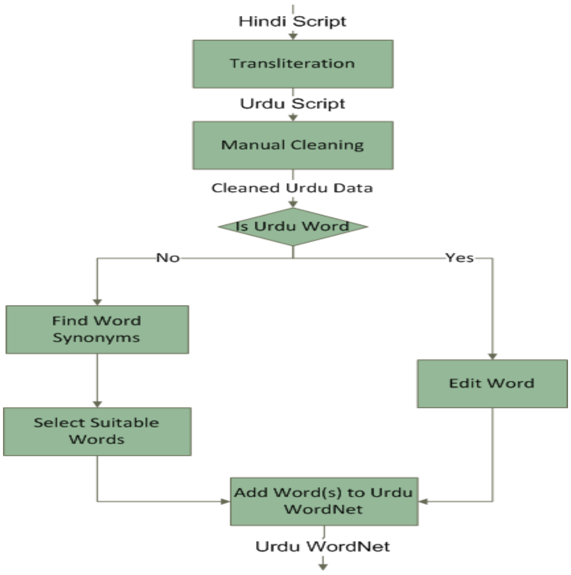

Figure 1: Hindi to Urdu WordNet Conversion

For the automatic transliteration we have developed the software which transliterates the Hindi script into Urdu script. For mapping of Hindi consonants, and vowels into Urdu number of rules are used depending on the position in the word i.e. same Hindi vowel would be mapped to different Urdu characters at the start, middle, and at the end of word e.g. ə is mapped to Alef (ا) + Zabar َ at the start of a word and by Zabar َ in the middle of a word. These rules are discussed in (Abbas Malik, 2008).

The transliteration system does not resolve the problem of multi-equivalences. For example the Hindi 'त' can be mapped to 'ت' *(Tay)* and 'ط'*(Tuay).* A list of multi-equivalence Hindi character is given in Table 2.

The multi-equivalence problem from Hindi to Urdu transliteration is problematic which needs to be solved. An automated method is applied to resolve this by analyzing the Urdu character frequency using an Urdu corpus i.e. to resolve the above problem the system map 'त' to 'ت' *(Tay)* due to more frequency of 'ت'*(Tay)* as compared to 'ط'*(Tuay)* .

Hindi	Urdu
अ	ا،آ*(Alif-Mad, Alif)*
त	ت،ط*(Tuay, Tay)*
स	ث،س،ص *(Suad, Seen, Say)*
ह	ح،ہ*(Hay)*
ज़	ز،ذ،ژ،ض،ظ*(Zueen, Zuad, Yeh, Zaal, Zeh)*
क	ق،ک*(Kaf)*

Table 2: Multiple Urdu characters for one Hindi character

Afterwards, multi-equivalence problem is resolved manually by analyzing the text. Although Hindi and Urdu are grammatically similar languages and share a large number of words, morphology, vocabulary and cultural heritage. But still there are number of Hindi words that are not used in Urdu. Therefore there is need to remove the Hindi words like انتیرن (*Anteeran)* (अंतीरन, fail) from Urdu WordNet. There are two steps to do this First find out the corresponding Urdu word and Second discard original Hindi Word. The deletion of Hindi word is shown in Figure 2.

ID :: 83
CAT :: adjective
CONCEPT :: जो परीक्षा में उत्तीर्ण न हुआ हो
EXAMPLE :: "रोहन परीक्षा में अनुत्तीर्ण हो गया"
SYNSET-HINDI :: अनुत्तीर्ण,फेल

Hindi WordNet Entry

ID :: 83
CAT :: adjective
CONCEPT :: جو پریکشا میں اُنتیرن نہ ہوا ہو
(Jo priksha mein Antteeran na huwa ho)
(Failed in exam)
EXAMPLE :: ""روہن پریکشا میں انتیرن ہوگیا
(Rohan priksha mein Antteeran ho gaya)
(Rohan has failed the test)
SYNSET-URDU :: انتیرن،فیل
(Anteeran , Fail)
(Fail)

Urdu WordNet Entry after Transliteration

ID :: 83
CAT :: adjective
CONCEPT :: امتحان میں ناکامیاب
(Imtihan mein nakamyab)
(Failed in exam)
EXAMPLE :: "روہن امتحان میں فیل ہوگیا"
(Rohan Imtihan mein fail ho gaya)
(Rohan has failed the test)
SYNSET-URDU :: فیل،ناکام
(Fail, Nakam)
(Fail)

Final Urdu WordNet Entry

Figure 2: Hindi Word to Urdu Word

Similarly, numbers of Urdu words are added in database, which are not present in Hindi WordNet. For example the word ربا (interest) is added in Urdu WordNet. The entry of ربا is shown in Figure 3.

ربا (Riba)
ID :: 7350
CAT :: noun
CONCEPT :: وہ رقم جو اصل پر زائد وصول کی جائے
(Woh raqam jo asal par zaid wasol ki jayey)
(The amount charged but more than the actual amount)
EXAMPLE :: ""اسلام میں ربا حرام ہے
(Islam mein Riba Haram Hai)
(Interest is forbidden in Islam)
SYNSET-URDU :: ربا،سود،بیاج
(Riba, Sood, Biyaj)
(Interest)

Figure 3: Sample New Word Added in Urdu WordNet

4 Urdu WordNet

The UWN currently has around 28967 synsets consisting of nouns, verbs, adverbs and adjectives. The detail of WordNet is shown in Table 3.

Category	Count
Noun	48224
Verb	3000
Adverb	705
Adjective	8000
Unique Words	50000
Synset	28967

Table 3: Urdu WordNet

Since it is currently in development phase so, new synset will get introduced in UWN. The front-end of the tool has been implemented in .NET. The application interface is connected at the backend with text files of synsets. The data is divided into 4 files i.e. Urdu-common, Urdu-core, Urdu-full and English. The synset entry format in file is shown in Figure 4.

ID: The synset identifier.
CAT: The syntactic category of the sense.
CONCEPT: It explains the concept represented by the synset.
EXAMPLE: It gives the usage of the words of the synsets in the sentence
SYNSET-URDU: It gives the set of synonyms for the sense in the Urdu language

Figure 4: Synset Entry Format

At present the offline version of Urdu WordNet is available which can be made available online after proper security implementation.

5 Discission & Future Work

This paper presents experience of building Urdu WordNet by Using the Hindi WordNet. The current Urdu WordNet does not provide the full-fledged lexical information of Urdu Words but, it can be used to extract the sense and synset information.

Although new Urdu words are added in Urdu WordNet which were missing in Hindi WordNet. Still there is need to add more Persian and Arabic load Urdu words to cover vocabulary of Urdu.

Diacritics are partially handled in Urdu WordNet. Currently there is no clear distinction between two words which have same written expression in case of no diacritic e.g. the بننا *(ban-na)(making)* and word بننا *(bun-na)(knitting)* are written as بننا *(ban-na)* in Urdu WordNet. The details of these words are given in Figure 4 & Figure 5. The Urdu WordNet system needs to be mature enough to handle diacritics. This can be achieved by adding up the diacritics in Urdu WordNet database.

ID :: 7132
CAT :: verb
CONCEPT :: روپ دینا
(Roop Dena)
(Form in a shape)
EXAMPLE :: "مندِربن گیا ہے"
(Mandir Bun gaya hai)
(The temple has constructed)
SYNSET-URDU :: بننا،تیار ہونا
(Ban-na , Tyar hona)

Figure 5: WordNet Entry for بننا (ban-na) (make)

ID :: 7310
CAT :: verb
CONCEPT :: ہاتھ یا اوزاروں سے کچھ سوتوں کو اوپر اُور کچھ کو نیچے سے نکال کر کوئی چیز بنانا
(Haath ya ozaron sey kuch soton ko ooper aur kuch ko nechey sy nikal ker koi cheez banana)
(Made something with the help of hand or tools by springs up and down)
EXAMPLE :: سیتا اپنے بچے کے ِلے ایک سویٹر بن رہی "ہے"
(Seeta Apney betey kay liyey saweeter bun rahi hai)
(Sitta is knitting a sweeter for her baby)
SYNSET-URDU :: بننا،بنائَی کرنا
(bun-na, bunaye karna)
(Knitting)

Figure 6: WordNet Entry for بننا (bun-na)(Knitting)

The semantic relations such like antonymy, hypernymy, hyponymy, me-ronymy, holonymy, troponymy, entailment etc. are ignored in Urdu WordNet. These relationships can be added to provide complete lexical information of Word.

The extension of Urdu WordNet further involves work in the area of compound words especially in the implementation of complex predicates e.g.

نکل گیا *(nikle gaya) (went out).* In Urdu 20% verb forms in the running text are compound verbs (Compound Verb, http://en.wikipedia.org/wiki/Compound_verb). So, there is need to add complex predicates which are used more frequently than normal verb.

Currently Compound words (Noun, adverbs) e.g. آہستہ آہستہ*(Ahista Ahista) (slowely Slowely)* are joined using "–"instead of Zero Width noun-joiner. There is need to add mechanism into WordNet tool to handle this issue.

6 Conclusion

In this paper, we present a report on development of Urdu WordNet by extracting information contained in existing Hindi WordNet. The scriptural barrier between two languages is crossed by using automatic and manual transliteration. Despite the similarity between two languages, concept translation is employed to remove Hindi words from Urdu WordNet. New Urdu words are also added in WordNet which are not present in Hindi WordNet.

Acknowledgements

We would like to thank Hindi WordNet Group at IIT Bombay for their support and especially Prof. Pushpak Bhattacharyya, Laxmi Kashyap, Salil Joshi, and Prabhakar Pandey. We acknowledge Asad Mustafa CLE-KICS for helping us in Urdu translation.

References

M. G. Abbas Malik , Christian Boitet , Pushpak Bhattacharyya, *Hindi Urdu machine transliteration using finite-state transducers*, Proceedings of the 22nd International Conference on Computational Linguistics, p.537-544, August 18-22, 2008, Manchester, United Kingdom

"Language Summary" Reterived July 2011 from, http://www.ethnologue.com/ethno_docs/distribution.asp?by=size

"Word Net Applications" Reterived July 2011 from, http://en.wikipedia.org/wiki/WordNet#Applications

Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande and P. Bhattacharyya, *An Experience in Building the Indo WordNet - a WordNet for Hindi*, First International Conference on Global WordNet, Mysore, India, January 2002.

"Hindi WordNet" Reterived July 2011 from http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php

"Hindi Word Net" Reterived July 2011 from, http://www.cfilt.iitb.ac.in/wordnet/webhwn/downloaderInfo.php

Hussain, Sarmad. 2004. *Letter-to-Sound Rules for Urdu Test to Speech System*, Proceeding of workshop on computational Approaches to Arabic Script-based Languages, COLING 2004, Geneva, Switzerland.

George A. Miller, Richard Beckwith, Christiane Fellbaum,Derek Gross, and Katherine J. Miller. 1993.*Five Papers on WordNet*. MIT press. http://www.mit.edu/~6.863/spring2009/readings/5papers.pdf

P.Vossen. 2002 Euro WordNet: General Document. University of Amesterdam

"Compound Verb" Reterived July 2011 from http://en.wikipedia.org/wiki/Compound_verb

Pushpak Bhattacharyya, *IndoWordNet*, Lexical Resources Engineering Conference 2010 (LREC 2010), Malta, May, 2010.

Debasri Chakrabarti, Vaijayanthi Sarma and Pushpak Bhattacharyya. 2007. *Complex Predicates in Indian Language* Wordnets, Lexical Resources and Evaluation Journal, 40 (3--4).

Tafseer Ahmed and Annette Hautli (2010). *Developing a Basic Lexical Resource for Urdu Using Hindi WordNet*. Proceedings of CLT10, Islamabad, Pakistan.

Alok Chakrabarty, Bipul Purkayastha and Arindam Roy. *Experiences In Building The Nepali Wordnet - Insights And Challenges*. The Fifth Global Wordnet Conference @ CFILT, IIT Bombay, Mumbai

Malhar Kulkarni, et al. (2010). *Introducing Sanskrit Wordnet*. The 5th International Conference of the Global WordNet Association (GWC-2010), 31st Jan - 4th Feb, 2010,