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 Nastaliq 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 الخ ط می ه لکھی ج اب (Urdu is written in Arabic script)

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

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>56623</td>
</tr>
<tr>
<td>Verbs</td>
<td>3894</td>
</tr>
<tr>
<td>Adjectives</td>
<td>13702</td>
</tr>
<tr>
<td>Adverbs</td>
<td>1276</td>
</tr>
<tr>
<td>Synsets</td>
<td>30977</td>
</tr>
</tbody>
</table>

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.cflit.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](image)

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. ऑ is mapped to Alef (ا) + Zabar (١) at the start of a word and by Zabar (١) 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 'Tay' (ت) and 'Ty(Tay)'. 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 maps 'त' (Tay) due to more frequency of 'ت' (Tuay) as compared to 'त' (Tuay).

Table 2: Multiple Urdu characters for one Hindi character

<table>
<thead>
<tr>
<th>Hindi</th>
<th>Urdu</th>
<th>Hindi WordNet Entry</th>
<th>Urdu WordNet Entry after Transliteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>अ</td>
<td>आ (Alif-Mad, Alif)</td>
<td>ID :: 83 CAT :: adjective CONCEPT :: ایک کام سے کامیاب (imtihaan mein nakamyab) EXAMPLE :: روہن میں انٹیرونز نے فیل بھیا (Rohan has failed the test) SYNSET-URDU :: انٹیرونز (Anteeran), فیل (Fail)</td>
<td>(Rohan failed the test)</td>
</tr>
<tr>
<td>त</td>
<td>त (Tuay, Tay)</td>
<td>ID :: 83 CAT :: adjective CONCEPT :: اس کو لیے (ruhne akele) EXAMPLE :: روہن میں انٹیرونز نے فیل بھیا (Rohan has failed the test) SYNSET-URDU :: انٹیرونز (Anteeran), فیل (Fail)</td>
<td></td>
</tr>
<tr>
<td>स</td>
<td>स (Suad, Seen, Say)</td>
<td>ID :: 83 CAT :: adjective CONCEPT :: اس کو لیے (ruhne akele) EXAMPLE :: روہن میں انٹیرونز نے فیل بھیا (Rohan has failed the test) SYNSET-URDU :: انٹیرونز (Anteeran), فیل (Fail)</td>
<td></td>
</tr>
<tr>
<td>ह</td>
<td>ह (Hay)</td>
<td>ID :: 83 CAT :: adjective CONCEPT :: اس کو لیے (ruhne akele) EXAMPLE :: روہن میں انٹیرونز نے فیل بھیا (Rohan has failed the test) SYNSET-URDU :: انٹیرونز (Anteeran), فیل (Fail)</td>
<td></td>
</tr>
<tr>
<td>ज</td>
<td>ज (Zueen, Zaad, Yeh, Zaal, Zeh)</td>
<td>ID :: 83 CAT :: adjective CONCEPT :: اس کو لیے (ruhne akele) EXAMPLE :: روہن میں انٹیرونز نے فیل بھیا (Rohan has failed the test) SYNSET-URDU :: انٹیرونز (Anteeran), فیل (Fail)</td>
<td></td>
</tr>
<tr>
<td>क</td>
<td>क (Kaf)</td>
<td>ID :: 83 CAT :: adjective CONCEPT :: اس کو لیے (ruhne akele) EXAMPLE :: روہن میں انٹیرونز نے فیل بھیا (Rohan has failed the test) SYNSET-URDU :: انٹیرونز (Anteeran), فیل (Fail)</td>
<td></td>
</tr>
</tbody>
</table>

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 अंतर (Anteeran) (अपमान, 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.

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

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.
Table 3: Urdu WordNet

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>48224</td>
</tr>
<tr>
<td>Verb</td>
<td>3000</td>
</tr>
<tr>
<td>Adverb</td>
<td>705</td>
</tr>
<tr>
<td>Adjective</td>
<td>8000</td>
</tr>
<tr>
<td>Unique Words</td>
<td>50000</td>
</tr>
<tr>
<td>Synset</td>
<td>28967</td>
</tr>
</tbody>
</table>

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

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

5 Discussion & 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 \( \text{ن ن ی ا} \) \((ban-na)\)\( \text{(making)} \) and word \( \text{ن ن ی ا} \) \((bun-na)\)\( \text{(knitting)} \) are written as \( \text{ن ن ی ا} \) \((ban-na)\) 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.

![Figure 4: Synset Entry Format](image)

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.
(nikle gaya) (went out). In Urdu 20% verb forms in the running text are compound verbs (Compound Verb, \url{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. آہ ست ي (Ahista Ahista) (slowly Slowly) 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


“Language Summary” Retreived July 2011 from, \url{http://www.ethnologue.com/ethno_docs/distribution.asp?by=size}


“Hindi WordNet” Retreived July 2011 from \url{http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php}

“Hindi Word Net” Retreived July 2011 from, \url{http://www.cfilt.iitb.ac.in/wordnet/webhwn/dow nloaderInfo.php}


“Compound Verb” Retreived July 2011 from \url{http://en.wikipedia.org/wiki/Compound_verb}


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