Developing Urdu WordNet Using the Merge Approach

Ayesha Zafar, Afia Mahmood, Farhat Abdullah, Saira Zahid, Sarmad Hussain, Asad Mustafa
Center for Language Engineering, Al-Khawarizmi Institute of Computer Science, University of Engineering and Technology, Lahore
firstname.lastname@kics.edu.pk

Abstract

The current paper describes the process of developing an Urdu WordNet. The process includes selecting words, identifying their senses and documenting their use. The current work also ties the Urdu senses with corresponding senses in English. Challenges in developing the WordNet and the solutions being implemented are discussed. Finally, this paper presents the work planned in the future.

1. Introduction

Fellbaum [1] defines WordNet as an extensive lexical database in which words are divided by part of speech and organized into a hierarchy of nodes. Each node represents a concept and words denoting the same concept are grouped into a synset with a unique ID, for example, ENG20-02853224-n: {car, auto, automobile, machine, motorcar}). Concepts are defined by a short gloss (e.g., 4-wheeled motor vehicle; usually propelled by an internal combustion engine) and are also linked to other relevant synsets in the database (e.g. hypernym: {motor vehicle, automotive vehicle}, hyponym: {cab, hack, taxi, taxicab}).

WordNet is used for many computational linguistic tasks such as Word Sense Disambiguation, Information Retrieval and Extraction and Machine Translation, etc. Over time, WordNet has become a valuable resource, which has initiated the development of WordNets for many other languages as well.

Urdu is a language of the Indo-Aryan family, widely spoken in Pakistan and India. It is written using Arabic script from right to left, in Nastalique writing style. Process for the development of Urdu WordNet has been discussed in this paper. The purpose of the development of Urdu WordNet is to provide a lexical resource for Urdu language that can be used in natural language processing. The WordNet is being developed specifically to align with linguistic, cultural, religious and other contexts in Pakistan.

The roadmap for the rest of paper is as follows: Section 2 presents the literature on Urdu WordNet. Methodology for development of Urdu WordNet is described in Section 3 and the current status is discussed in Section 4. Section 5 discuses the relevant issues and solutions, and Section 6 concludes the paper.

2. Literature Review

WordNets in various languages have been developed both through manual [2, 13] and automated [3, 14] methods. The manual construction of each WordNet is more accurate, but is also more time-consuming and expensive. There are two common approaches for building a WordNet for a language [4]: (i) a top-down approach, using an existing WordNet in a source language to seed the linguistic data for the target language WordNet [4], and (ii) a bottom-up approach, where the linguists create the WordNet synsets without depending on an existing one [5].

In the top-down approach, the synsets from the source language are translated into the target language. However, for the synsets to be mappable, concepts in the source language must exist in the target language, which is not always possible. Additionally, generally a significant amount of language resource is required for building a WordNet. For example, a set of synsets strictly aligned with the source WordNet must exist before the new WordNet can be built. This is a significant drawback of building a WordNet from an existing one. For this approach to be

successful there must be significant level of linguistic similarity between the two languages [5, 6].

Two methods have been discussed for developing a WordNet through the bottom-up process: the merge approach and the expand approach [7]. The merge approach builds the taxonomies of the language, synsets and relations, and then map to the Princeton WordNet (PWN) by using the English equivalent words from existing bilingual dictionaries [15]. Merge approach provides a description of lexico-semantic relations, closer to the spirit of the given language, in that it is less influenced by the design decisions in a WordNet for another language, often of a significantly different type. The merge approach, however, requires rich resources at the outset, for example, a monolingual dictionary with senses identified, detailed definitions, thematic codes for senses and some semantic structuring [15].

The expand approach is to map or translate local words directly to the PWN's synsets by using the existing bilingual dictionaries. Thai WordNet construction has used the expand approach due to budget and time reasons [7].

Previous work on Urdu WordNet [8, 9] is based on the top-down approach. Hindi WordNet (HWN) has been used due to its similarity with Urdu. However, this method faced the following challenges [8].

- There are number of Hindi words that are not used in Urdu due to the linguistic, religious, cultural and other differences, e.g. انتیرن (fail) is not normally used in Urdu.
- Many words which are commonly used in Urdu, e.g. those loaned from Arabic and Persian languages, are not present in Hindi WordNet synsets. For example ربا (interest) is used in Urdu but not available in HWN.
- In the explanation given for the synset and the example for its usage a lot of Hindi words are used, which are not part of the common cultural vocabulary of Urdu in Pakistan. For example in the sentence. ارجن پریکشا میں کامیاب رہا. ارجن and پریکشا are not commonly used.

  In addition, the compound words and complex predicates in verbs are not addressed.

3. Methodology

To build Urdu language WordNet merge approach has been used. 5000 high frequency nouns, verbs, adjectives and adverbs are selected from Urdu corpus [10] to develop the WordNet. The following process is used for the development of Urdu WordNet.

1. A word from the list of 5000 words is looked up into Urdu Lughat [11]
2. Its POS tag is determined by Urdu Lughat. For example the word کھانا which has two POS tags in Urdu Lughat i.e. کھانا (meal) a noun and کھانا (eat) a verb.
3. The number of senses for each POS of the particular word is determined from Urdu Lughat. The less common, literary and poetic senses are ignored. So the number of senses for each word varies according to its use. For example, the third sense is in Table 1 below is less common and poetic, and thus ignored.

Table 1: Urdu Word Senses

Concept	Sense	English Translation
پکڑا ہوا، قیدی، محبوس	گرفتار	Capture
مبتلا، پھنسا ہوا، گھرا ہوا	گرفتار	Entangled
عاشق، فریفتہ	گرفتار	Smitten

4. The English translation of the word according to its POS tag is looked up in Urdu to English Dictionary. If there are two or more POS tags of the word in Urdu Lughat then the English translation of the word is determined according to all its tags as the word کھانا (meal) is a noun as well as a verb کھانا (eat) . So both the categories will be created. Figure 1 shows different POS categories of the word کھانا.

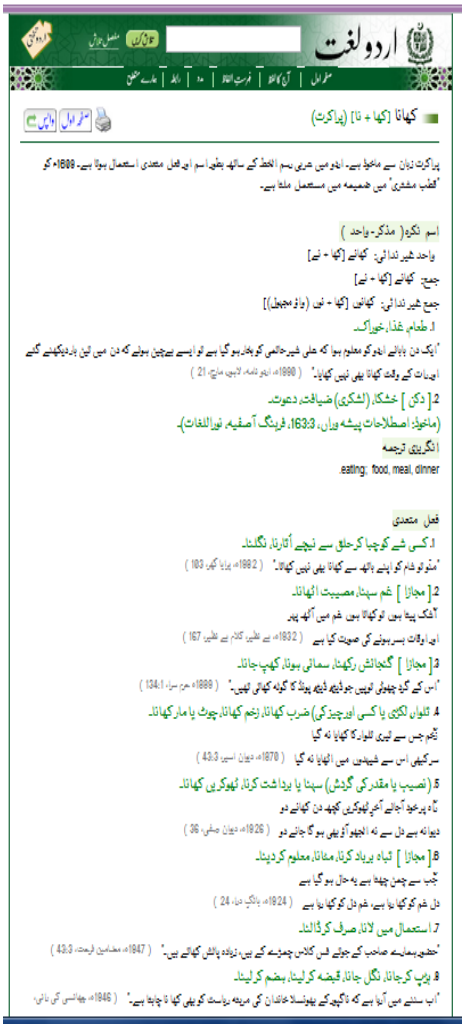

Figure 1: POS Cat. of کھانا in Urdu Lughat

5. English translation of an Urdu word may be different for its multiple senses. So the English translation of each sense is looked up separately in Urdu to English Dictionary. The example is explained in the Table 2.

Table 2: English Translation of Urdu Word

English word	Concept of each sense	Urdu Word
Work	کسبِ معیشت کا وسیلہ یا ذریعہ	کام
Chores	روزمرہ یا مقررہ وقت کا کام	کام
Concern	سروکار یا واسطہ ہونا	کام
embroidery	کڑھائی، نقاشی وغیرہ کا کام	کام

6. The selected word is looked up in Princeton WordNet version 2.1 and each sense of Urdu is mapped on the sense of English according to its determined POS tag. The unique ID of English sense and its English word is recorded in separate columns. Table 3 shows the unique ID of English sense.

Table 3: Unique IDs of English Senses

English ID	English Word
578942	Work
708623	Chores
5600606	Concern
3248411	Embroidery

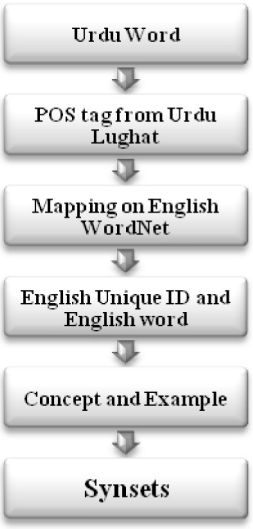

Figure 2: Urdu WordNet Process

7. The concept of each sense is explained with the help of Urdu Lughat in simple and precise language.
8. Further, an example is given to illustrate the concept, using a word from the synset. For formulating the example, as a first preference the example usage given Urdu Lughat is used. If this example is difficult to understand, a new example sentence is created. Where it is not easily possible, the corresponding example from PWN is translated as an alternative.
9. The synsets of the word are written from Qamos-e-Mutradifat (synonyms dictionary) [12]. Only those synonyms from Qamos-e-Mutradifat are selected that have the same concept. The concepts of these synonyms are confirmed from Urdu Lughat.

10. In the end, a linguist reviews the WordNet entries.

This process is summarized in the Figure 2.

4. Current Status

A sample Urdu WordNet entry is given in the table below.

Table 4: Urdu WordNet Entries

Synsets	دباؤ، بوجھ، بھار، بار، ثقل	دباؤ، سختی، جبر	دباؤ، خوف، ڈر، دہشت
Urdu ID[1]	1	2	3
Category	N	N	N
Concept	کسی چیز کا بوجھ یا وزن	سختی یا جبر کرنے کا عمل	خوف یا دہشت ہونا
Example	اس نے میز پر دباؤ ڈالا تو وہ ٹوٹ گئی	بچوں پر غیر ان ضروری دباؤ کو والدین سے متنفر کر دیتا ہے	وڈیرے کے دباؤ کی وجہ سے وہ کچھ نہ بولا
English ID	11329024	416551	7418507
English Word	Pressure	Oppression	terror

At present, 2205 senses are completed. These include 1518 nouns, 560 adjectives, 80 verbs and 47 adverbs.

5. Discussion

This paper presents experience of building Urdu WordNet. Although it gives sufficient lexical information of Urdu words but still there are issues needed to be resolved. Some language specific challenges are observed during the development of Urdu WordNet process that are needed to be considered carefully. The diacritics need to be handled for Urdu. The words that change their meaning with the diacritics need to have a separate entry in Urdu WordNet. Table 5 shows the example. This is addressed in the Urdu WordNet.

Table 5: The Case of Diacritics

Urdu	**Concept**	**English**
گنّا	بانس کے درخت کی وضع کا پودا جو	sugar cane
گِننا	گننا، شمار کرنا	count

There are Urdu words/concepts that do not exist in the English WordNet due to religious, cultural and other differences. Some examples are given in Table 6.

Table 6: The Case of Cultural Concepts

Words	**Concept**
صفر	name of the second Islamic month
مہندی	a cultural function which is celebrated before the marriage ceremony in which typical intricate patterns of Henna are applied to bride, celebrated mainly by the bride's family
ڈوپٹہ	a long scarf that is worn by females to cover their head

This difference creates problem when Urdu synset is mapped onto English ID.

Further, because of the difference in the structure of English and Urdu language it is difficult to map some of the words on the same POS tag. For example the word قیدی "prisoner" is a noun in English but Urdu Lughat lists it as an adjective. صارف "consumer" is a noun in English and an adjective in Urdu. Similarly the word پولنگ "polling" is a noun in Urdu and a verb in English. In order to incorporate this problem, there is need to improve Urdu Lughat.

Sometimes two different words are mapped on the same English ID, to avoid this problem and keep all the IDs unique that particular word is added into the synset of the previously added word.

In the future, 5000 senses will be completed. Currently nouns are more in number than other categories. The words added in the future will be

[1] This is an arbitrarily assigned number, which will be finalized upon release of Urdu WordNet.

selected from other categories as much as possible, to balance this distribution. Further the work will associate these synsets, to allow for more significant modeling of the semantic relationships.

6. Conclusion

In this paper, we present the process of developing a basic lexical resource for Urdu. This lexical resource is developed using the bottom-up approach. A few language and cultural issues faced in its development are discussed. This is a work in progress and future goals are also presented.

7. Acknowledgements

This work has been supported through a grant from DAAD Germany, and is being done in collaboration with University of Konstanz, Germany.

References

[1] C. Fellbaum, "WordNet: An Electronic Lexical Database." MIT Press, Cambridge, Massachusetts, 1998.

[2] C. Fellbaum, M. Palmer, L. Delfs, S. Wolf, "Manual and Automatic Semantic Annotation with WordNet", 2001, Retrieved (06, 27, 2012). Available at: https://nats-www.informatik.uni-hamburg.de/intern/proceedings/2001/naacl/mwnw/pdf/invitedPaper.pdf

[3] M. Montezory and Heshaam Faili "Automatic Persian WordNet Construction" Coling 2010: Poster Volume, pages846–850, Available at: http://aclweb.org/anthology/C/C10/C10-2097.pdf

[4] M. Khan and F. Faruqe. "*BWN- A Software Platform for Developing Bengali WordNet*", Center for research on Bangla language processing (CRBLP), BRAC University 2010, Retrieved (06, 27, 2012). Available at: http://crblp.bracu.ac.bd/papers/2008/BWN-architecture-CISSE08.pdf

[5] D. Fiser, "A Multilingual Approach to Building Slovene WordNet" In Proc. *Workshop on A Common Natural Language Processing Paradigm for Balkan Languages held within the Recent Advances in Natural Language Processing Conference RANLP'07*. Bulgaria, 2007.

[6] E. Barbu and V. B. Mititelu, "Automatic Building of WordNets" In Proc. *International Conference Recent Advances in Natural Language Processing*, Borovets, Bulgaria, 2005, Retrieved (06, 27, 2012). Available at: http://clic.cimec.unitn.it/eduard/publications/AutomaticBuildingWordnets.pdf

[7] S. Thoongsup, K. Robkop, C. Mokarat, T. Sinthurahat, T. Charoenporn, V. Sornlertlamvanich and H. Isahara, "Thai WordNet Construction" In. Proc. *7th Workshop on Asian Language Resources, ACL-IJCNLP*, 2009, pages 139–144, Retrieved (06, 27, 2012). Available at: http://aclweb.org/anthology/W/W09/W09-3420.pdf

[8] F. Adeeba and S. Hussain, "Experiences in Building the Urdu WordNet", *IJCNLP*, 2011, Retrieved (06, 27, 2012). Available at: http://www.cle.org.pk/Publication/papers/2011/UrduWordNet.pdf

[9] T. Ahmed and A. Hautli, "Developing a Basic Lexical Resource for Urdu using Hindi WordNet", in proc. *CLT10*, Islamabad, 2010, Retrieved (06, 27, 2012). Available at: http://ling.uni-konstanz.de/pages/home/pargram_urdu/main/files/Ahmed_Hautli_CLT10.pdf

[10] M. Ijaz and S. Hussain, "*Corpus Based Urdu Lexicon Development*", Centre for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Lahore, 2007, Retrieved (06, 27, 2012). Available at: http://cle.org.pk/Publication/papers/2007/corpus_based_urdu_lexicon_development.pdf

[11] *Urdu Lughat*, Urdu Lughat Board Karachi, 2002, Retrieved (06, 27, 2012). Available at: http://www.clepk.org/oud/Default.aspx

[12] W. Sirhindi, *Qamoos-e-Mutradifaat*, Urdu Science Board, Lahore, 2006.

[13] P. Sathapornrungkij, C. Pluempitiwiriyawej, "*Construction of Thai WordNet Lexical Database from Machine Readable Dictionaries*", Mahidol University, Bangkok, 2005, Retrieved (06, 27, 2012). Available at: http://www.mt-archive.info/MTS-2005-Sathapornrungkij.pdf

[14] M. Saveski, I, Trajkovski, "*Automatic Construction of Wordnets by Using Machine Translation and Language Modeling*", Staffordshire University, UK, 2010, Retrieved (06, 27, 2012). Available at: http://www.time.mk/trajkovski/papers/is2010.pdf

[15] M. Piasecki, S. Szpakowicz, B. Broda *"A Wordnet from the Ground Up"*, Oficyna Wydawnicza Politechniki Wroclawskiej, Wroclaw, 2009, Retrieved (24, 8, 2012). Available at: http://www.site.uottawa.ca/~szpak/pub/A_Wordnet_from_the_Ground_Up.pdf