Developing Urdu Wordnet using the Merge Approach

Afia Mahmood

INTRODUCTION

- WordNet is a large lexical database in which nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synsets, each expressing a distinct concept.
- Synsets are interlinked by means of conceptual-semantic and lexical relations.
- The resulting network of meaningfully related words and concepts can be navigated with the browser . (http://wordnet.princeton.edu/)

Contents in Wordnet

- Code
- POS Category
- Concept
- Example
- Synsets
- Semantic relations

Example:

- {02791984} <noun.artifact> bed#1 -- (a piece of furniture that provides a place to sleep; "he sat on the edge of the bed"; "the room had only a bed and chair")
- {02792607} <noun.artifact> bed#2 -- (a plot of ground in which plants are growing; "the gardener planted a bed of roses")
- {09085476} <noun.object> bed#3, bottom#5 -- (a depression forming the ground under a body of water; "he searched for treasure on the ocean bed")

USAGE

- WordNet is used for many computational linguistic tasks such as Word Sense Disambiguation,
- Information Retrieval and Extraction
- and Machine Translation, etc.

URDU WORDNET

- 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.

- To build Urdu language WordNet merge approach has been used

WHAT IS MERGE APPROACH?

- 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.

METHODOLOGY

5000 high frequency words including:

- nouns,
- verbs,
- adjectives
- and adverbs

are selected from Urdu corpus to develop the WordNet.

PROCESS:

1. A word from the list of 5000 words is looked up into Urdu Lughat

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** OR **poetic** senses are ignored. So the number of senses for each word varies according to its use.

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.

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.

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.
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 given in 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) 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.

Sample Urdu Worndet sheet

CHALLENGES

1. 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.

Urdu Word	Concept	English
گنا	بانس کے درخت کی وضع کا پودا جو رسیلا اور میٹھا ہوتا ہے	sugar cane
گِنا	گننا، شمار کرنا	count

- 2. There are Urdu words/concepts that do not exist in the English WordNet due to religious, cultural and other differences.

Words	Concept	Words
صفر	name of the second Islamic month	صفر
مہندی	a cultural function which is celebrated one day before the marriage ceremony in which typical intricate patterns of Henna (paste of myrtle leaves) are applied to bride. It is a fun filled ritual, which is celebrated mainly by the bride's family	مہندی

- 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 incorporate these words and tags in 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.

THANK YOU!