pyiwn: A Python-based API to access Indian Language WordNets

Ritesh Panjwani, Diptesh Kanojia, Pushpak Bhattacharyya
{riteshp, diptesh, pb}@cse.iitb.ac.in

Center For Indian Language Technology,
Indian Institute of Technology Bombay, India

At GWC 2018, Singapore, January 2018

Motivation

- Efforts to create a lexical semantic network for Indian languages began with Hindi Wordnet (Narayan et al., 2002)
- Based on the concept of pivotal expansion, IndoWordnet was created (Bhattacharyya, 2010)
- IndoWordnet is a linked structure of Wordnets of major Indian languages from Indo-Aryan, Dravidian and Sino-Tibetan families
- The API bundled with the IndoWordnet data could be really helpful to the NLP community

Introduction

- The API provides access to the synsets and its lexical and semantic relations for 18 major Indian languages
- Getting started with the pyiwn API:
  - It can be installed using `pip`:
    ```
    >>> pip install pyiwn
    ```
  - It can be imported like this:
    ```
    >>> from pyiwn import pyiwn
    ```
  - Downloading the IndoWordnet synset data
    ```
    >>> pyiwn.download()
    ```
  - Choosing a language to access its Wordnet
    ```
    >>> iwn = pyiwn.IndoWordNet('hindi')
    ```

Access to Synsets

- **All Synsets**: All the synsets and words for the given language can be accessed together with an optional `pos` argument which lets you constrain the part of speech of the word:
  ```
  >>> iwn.all_synsets(pos=pyiwn.ADVERB)
  >>> iwn.all_words(pos=pyiwn.NOUN)
  ```
- **Specific Synsets:** The specific synsets for a given word can be accessed with an optional `pos` argument which lets you constrain the part of speech of the word:
  ```
  >>> synsets=iwn.synsets('सज्जन', pos=pyiwn.NOUN)
  [Synset('सज्जन.noun.221')]
  ```
- **Synset Properties:** The synset properties like, head word, POS tag, gloss, examples, lemmas can be accessed like this:
  ```
  >>> synset=synsets[0]
  >>> synset.pos()
  >>> synset.lemmas()
  >>> synset.gloss()
  >>> synset.examples()
  ```

Access to Synset relations

- **Semantic relations**
  - The synset relations like hypernymy, function verb, modifies verb, modifies noun, ability verb, *etc*. can be accessed like this:
    ```
    >>> synsets=iwn.synsets('सज्जन', pos=pyiwn.NOUN)
    >>> synset=synsets[0]
    >>> synset.hypernymy()
    [Synset('देवालय.noun.451')]
    ```
  - The other semantic relations can also be accessed in a similar fashion
- **Lexical relations**
  - Some relations like antonymy that defined over lemmas can be accessed like this:
    ```
    >>> synset=iwn.synsets('सुबह', pos=pyiwn.NOUN)[0]
    >>> lemma = synset.lemmas()[0]
    Lemma('सुबह.noun.26824.सुबह')
    >>> lemma.antonym()
    [Synset('शाम.noun.8164')]
    ```
- The complete list of relations can be found in the documentation (refer the URL or QR code below)

Other features

- **Morphological analyzers**: The API also provides Morphological analyzers for Hindi and Marathi languages
  ```
  >>> iwn.morph('किसानों')
  # return the dictionary form (lemma): किसान
  ```
- **Speech data**: The speech data for words in Hindi Wordnet can also be accessed via the API using the following function:
  ```
  >>> iwn.speech('किसान')
  # returns a WAV file object
  ```

Conclusion and future work

- We provide an API for accessing Indian language Wordnets in the IndoWordnet using Python
- In future, we plan to add functionalities like:
  - Getting the top-level relational synset, the path-length of longest and shortest relational synset, similarity measures, *etc*.
  - Morphological analyzers for more languages other than Hindi and Marathi
  - Speech data for more languages

References:

- Christiane Fellbaum. 1998. WordNet. Wiley Online Library.
- Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande, and Pushpak Bhattacharyya. 2002. An experience in building the indowordnet-a wordnet for hindi. In First International Conference on Global WordNet, Mysore, India.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural language processing with Python: analyzing text with the natural language toolkit. ” O’Reilly Media, Inc.”.
- P Bhattacharyya. 2010. Indowordnet. Lexical Resources Engineering Conference 2010 (LREC 2010). Malta, May.

API URL: https://github.com/riteshpanjwani/pyiwn

Scan the QR code to access the API