IndoWordNet::Similarity
Computing Semantic Similarity and Relatedness using IndoWordNet

Sudha Bhingardive, Hanumant Redkar, Prateek Sappadla, Dhirendra Singh, Pushpak Bhattacharyya
Center for Indian Language Technologies,
Indian Institute of Technology Bombay, India
{bhingardivesudha, hanumantredkar, prateek2693, dhiru.research, pushpakbh}@gmail.com

Abstract

Semantic similarity and relatedness measures play an important role in natural language processing applications. In this paper, we present the IndoWordNet::Similarity tool and interface, designed for computing the semantic similarity and relatedness between two words in IndoWordNet. A java based tool and a web interface have been developed to compute this semantic similarity and relatedness. Also, Java API has been developed for this purpose. This tool, web interface and the API are made available for the research purpose.

1 Introduction

The Semantic Similarity is defined as a concept whereby a set of words are assigned a metric based on the likeliness of the semantic content. It is easy for humans with their cognitive abilities to judge the semantic similarity between two given words or concepts. For example, a human can quite easily say that the words *apple* and *mango* are more similar than the words *apple* and *car*. There is some understanding of how humans are able to perform this task of assigning similarities. However, measuring similarity computationally is a challenging task and attracts a considerable amount of research interest over the years. Another term very closely related to similarity is Semantic Relatedness. For example, *money* and *bank* would seem to be more closely related than *money* and *cash*. In past, various measures of similarity and relatedness have been proposed. These measures are developed based on the lexical structure of the WordNet, statistical information derived from the corpora or a combination of both. These measures are now widely used in various natural language processing applications such as Word Sense Disambiguation, Information Retrieval, Information Extraction, Question Answering, *etc.*

We developed IndoWordNet::Similarity tool, interface and API for computing the semantic similarity or relatedness for the Indian Languages using IndoWordNet.

The paper is organized as follows. Section 2 describes the IndoWordNet. Semantic similarity and relatedness measures are discussed in section 3. Section 4 details the IndoWordNet::Similarity. Related work is presented in section 5. Section 6 concludes the paper and points to the future work.

2 IndoWordNet

WordNet[1] is a lexical resource composed of synsets and semantic relations. Synset is a set of synonyms representing distinct concept. Synsets are linked with basic semantic relations like hypernymy, hyponymy, meronymy, holonymy, troponymy, *etc*. and lexical relations like antonymy, gradation, *etc.* IndoWordNet (Bhattacharyya, 2010) is the multilingual WordNet for Indian languages. It includes eighteen Indian languages *viz.*, *Assamese*, *Bengali*, *Bodo*, *Gujarati*, *Hindi*, *Kannada*, *Kashmiri*, *Konkani*, *Malayalam*, *Manipuri*, *Marathi*, *Nepali*, *Odia*, *Punjabi*, *Sanskrit*, *Tamil*, *Telugu*, *Urdu*, *etc*. Initially, Hindi WordNet[2] was created manually taking reference from Princeton WordNet. Similarly, other Indian language Word-

[1] http://wordnet.princeton.edu/
[2] http://www.cfilt.iitb.ac.in/indowordnet/

Nets were created from Hindi WordNet using expansion approach and following the three principles of synset creation. In this paper, we present the IndoWordNet::Similarity tool, interface and API, which help in computing similarity and relatedness of words / concepts in Indian language WordNets.

3 Overview of Semantic Similarity and Relatedness Measures

Over the years, various semantic similarity and relatedness measures have been proposed. These measures are classified based on the path length, information content and the gloss overlap. Some of them are described below.

3.1 Path Length Based Measure

These measures are based on the length of the path linking two synsets and the position of synset the WordNet taxonomy.

3.1.1 Shortest Path Length Measure

This is the most intuitive way of measuring the similarity between two synsets. It calculates the semantic similarity between a pair of synsets depending on the number of links existing between them in the WordNet taxonomy. The shorter the length of the path between them, the more related they are. The inverse relation between the length of the path and similarity can be characterized as follows:

$$sim_{path} = \frac{1}{shortest\_path\_length(S_1,S_2)}$$

$$sim_{path} = 2 * D - shortest\_path\_length(S_1,S_2)$$

Where, S_1 *and* S_2 are synsets and *D* is the maximum depth of the taxonomy.

3.1.2 Leacock and Chodorow's Measure

This measure proposed by Leacock and Chodorow's (1998) computes the length of the shortest path between two synsets and scales it by the depth *D* of the *IS-A* hierarchy.

$$sim_{lch} = -\log(\frac{shortest\_path\_length(S_1,S_2)}{2 * D})$$

Where, S_1 and S_2 are the synsets and *D* represents the maximum depth of the taxonomy.

3.1.3 Wu and Palmer Measures

This measure proposed by Wu & Palmer (1994) calculates the similarity by considering the depths of the two synsets, along with the depth of the lowest common subsumer (LCS). The formula is given as,

$$sim_{wup} = \frac{2 * depth(LCS(S_1,S_2))}{depth(S_1) + depth(S_2)}$$

Where, S_1 and S_2 are the synsets and $LCS(S_1,S_2)$ represents the lowest common subsumer of S_1 and S_2.

3.2 Information Content Based Measure

These measures are based on the information content of the synsets. Information content of a synset measures the specificity or the generality of that synset, *i.e.* how specific to a topic the synset is.

3.2.1 Resnik's Measure

Resnik (1995) defines the semantic similarity of two synsets as the amount of information they share in common. It is given as,

$$sim_{resnik} = IC(LCS(S_1,S_2))$$

This measure depends completely upon the information content of the lowest common subsumer of the two synsets whose relatedness we wish to measure.

3.2.2 Jiang and Conrath's Measure

A measure introduced by Jiang and Conrath (1997) addresses the limitations of the Resnik measure. It incorporates the information content of the two synsets, along with that of their lowest common subsumer. This measure is given by the formula:

$$distance_{jcn}(S_1,S_2) = IC(S_1) + IC(S_2) - (2\,IC(LCS(S_1,S_2)))$$

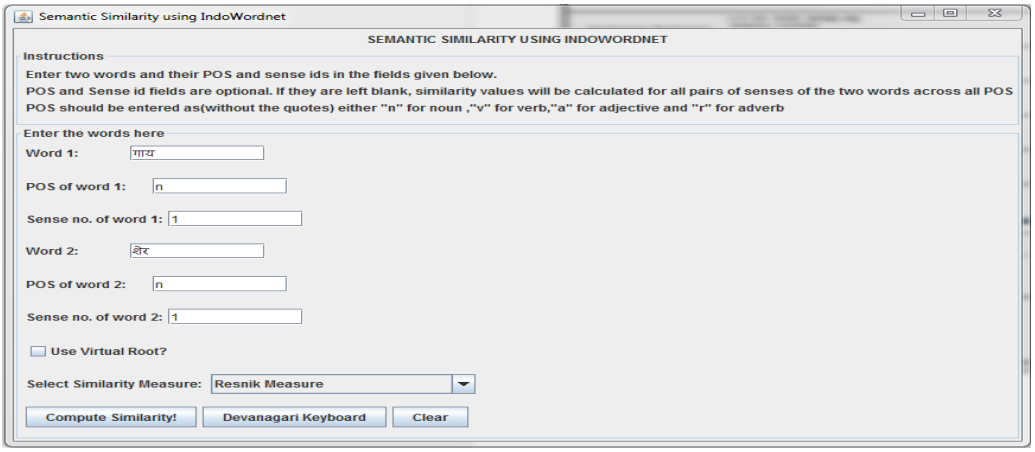

Figure 1: IndoWordNet::Similarity Tool

Where, *IC* determines the information content of a synset and *LCS* determines the lowest common subsuming concept of two given concepts.

3.3 Gloss Overlap Measures

Lesk (1986) defines the relatedness in terms of dictionary definition overlap of given synsets. Further, the extended Lesk measure (Banerjee and Pedersen, 2003) computes the relatedness between two synsets by considering their own gloss as well as by considering the gloss of their related synsets.

4 IndoWordNet::Similarity

We have developed IndoWordNet::Similarity tool, web based interface and API to measure the semantic similarity and relatedness for a pair of words / synsets in the IndoWordNet.

4.1 IndoWordNet::Similarity Tool

The IndoWordNet::Similarity[3] tool is implemented using Java. The user interface layout and its features are given below.

4.1.1 User Interface Layout

The main window of the tool is as shown in Figure 1. In order to use this tool, user needs to provide the following inputs:

- User can enter the pair of words for which similarity to be computed.
- User can specify the part-of-speech and the sense number for the given two words for calculating the similarity. If user doesn't provide these details then the tool computes the similarity between all possible pair of senses of the two input words over all parts-of-speech.
- Drop-box is provided for selecting the type of similarity measure.
- Check-box is provided for virtual root option.

Depending on the user query the similarity is calculated and displayed in an output window.

4.1.2 Features

- This is system independent portable standalone Java Application.
- Option such as part-of-speech and sense-id are optional.
- If user doesn't provide part-of-speech and sense-id option, then similarity is calculated for all possible pair of senses of the given words.
- If the virtual root node option is enabled then one hypothetical root is created which connects all roots of the taxonomy. This allows similarity values to be calculated between any pair of nouns or verbs.

4.2 IndoWordNet::Similarity API

IndoWordNet::Similarity Application Programming Interface (API) has been developed using Java which provides functions to compute the semantic similarity and relatedness using various measures. API provides three types of functions for each measure.

1. A function which takes only two words as

[3] http://www.cfilt.iitb.ac.in/iwnsimilarity

parameters and returns the similarity score between all possible senses of the two words.

2. A function which takes two words along with part-of-speech, sense-id and returns the similarity score between the particular senses as specified by the user.
3. A function which takes only two words as parameters and returns the maximum similarity between two words among all possible sense pairs. Some of the API functions are mentioned below:

API Function	Computes
public SimilarityValue[] getPathSimilarity(String word1, String pos1, int sid1, String word2, String pos2, int sid2, boolean use_virtual_root)	Path Similarity
public SimilarityValue[] getPathSimilarity(String word1,String word2,boolean use_virtual_root)	Path Similarity
public SimilarityValue getMaxPathSimilarity(String word1, String word2, boolean use_virtual_root)	Maximum Path Similarity

Table 1. Important functions of IndoWordNet::Similarity API

4.3 IndoWordNet::Similarity Web Interface

IndoWordNet::Similarity Web Interface has been developed using Php and MySql which provides a simple interface to compute the semantic similarity and relatedness using various measures. Figure 2 shows the IndoWordNet::Similarity web interface.

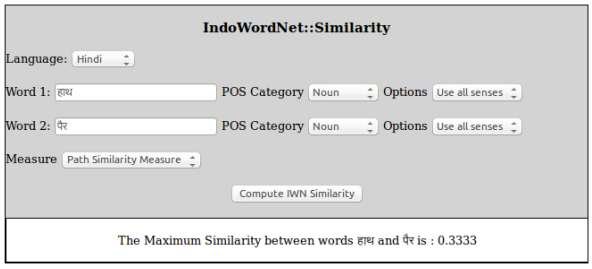

Figure 2. IndoWordNet::Similarity Web Interface

5 Related Work

WordNet::Similarity[4] (Pedersen *et. al*. 2004) is freely available software for measuring the semantic similarity and relatedness for English WordNet. This application uses an open source Perl module for measuring the semantic distance between words. It provides various semantic similarity and relatedness measures using WordNets. Given two synsets, it returns numeric score showing their degree of similarity or relatedness according to the various measures that all rely on WordNet in different ways. It also provides support for estimating the information content values from untagged corpora, including plain text, the Penn Treebank, or the British National Corpus[5].

WS4J[6] (WordNet Similarity for Java) provides a pure Java API for several published semantic similarity and relatedness algorithms. WordNet Similarity is also integrated in NLTK tool[7]. However, the need to make entirely different application for IndoWordNet lies in its multilingual nature which supports 19 Indian language WordNets. Hence, we developed the IndoWordNet::Similarity tool, web interface and API for calculating the similarity and relatedness.

6 Conclusion

We have developed the IndoWordNet::Similarity tool, web interface for computing the semantic similarity and relatedness measures for the IndoWordNet. Also, a java API has also been developed for accessing the similarity measures. The tool and the API can be used in various NLP areas such as Word Sense Disambiguation, Information Retrieval, Information Extraction, Question Answering, *etc*. In future, the other measures of computing similarity and relatedness shall be integrated in our tools and utilities.

References

Satanjeev Banerjee and Ted Pedersen. 2003. *Extended gloss overlaps as a measure of semantic relatedness*. In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, pages 805–810, Acapulco, August.

[4] http://wn-similarity.sourceforge.net/
[5] http://corpus.byu.edu/bnc/
[6] https://code.google.com/p/ws4j/
[7] http://www.nltk.org/howto/wordnet.html

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

Jay Jiang and David Conrath. 1997. *Semantic similarity based on corpus statistics and lexical taxonomy*. In Proceedings on International Conference on Research in Computational Linguistics, pages 19–33, Taiwan.

Claudia Leacock and Martin Chodorow. 1998. *Combining local context and WordNet similarity for word sense identification*. In C. Fellbaum, editor, WordNet: An electronic lexical database, pages 265–283. MIT Press.

Michael Lesk. 1986. *Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone*. In Proceedings of SIGDOC '86, 1986.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. *Wordnet::Similarity - Measuring the relatedness of concepts*. In Proceedings of AAAI04, Intelligent Systems Demonstration, San Jose, CA, July 2004.

Philip Resnik. 1995. *Using information content to evaluate semantic similarity in a taxonomy*. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pages 448–453, Montreal, August.

Zibiao Wu and Martha Palmer. 1994. *Verb semantics and lexical selection*. ACL, New Mexico.