MULTILINGUAL DICTIONARY GENERATION USING INDO-WORDNET: A PROPOSAL

Aadil A. Lawaye*
Neha Dixit**

Abstract

This paper attempts to automatically build a multilingual dictionary using Indo-WordNet. WordNet is a lexical database for different languages. It groups words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. For Indian languages, an Indo-WordNet is developed which includes languages like Hindi, Bengali, Marathi, Punjabi, Urdu, Kashmiri, Tamil, Kannada, Telugu, Malayalam, Assamese and Oriya.

Key words: Indo-Word Net, Multilingual, Dictionary.

Introduction

Dictionary describes the meaning of words, often illustrating how they are pronounced and used in context. Modern dictionaries often include information about spelling, etymology, usage, synonyms and grammar, and some may include illustrations as well. In many languages, words can appear in many different forms, but only the un-declined or un-conjugated form appears as the head word in most dictionaries, rather we can say that the words are looked at from a lexeme perspective. Dictionaries can vary widely in coverage, size and scope. A maximizing dictionary lists as many words as possible from a particular speech community, whereas minimizing dictionary exclusively attempts to cover only a limited selection of words from a speech community. Corpus-based dictionary is to provide learners with relevant, idiomatic and useful information that will help them setting up native-like links between words and meanings. In a corpus based dictionary lexicographers are keen to include corpus information about lexico-semantic relations such as synonyms, antonyms, hyponyms and super ordinates.

Natural language is inherently ambiguous. A word can have multiple meanings (or senses). Given an occurrence of a word *w* in a natural language text, task of Word Sense Disambiguation (WSD) is to determine the correct sense of word in that context. WSD is a fundamental and central open problem of Natural Language Processing (NLP). Highly ambiguous words pose continuing problems for NLP applications. They can lead to irrelevant document retrieval translations in Machine Translation systems (Palmer et al., 2000). Lexical ambiguity is syntactic or semantic. A word's syntactic ambiguity can be resolved by applying part-of-speech taggers which predict the syntactic category of a
word in texts with high levels of accuracy (Brill, 1995; Brants, 2000). The problem of resolving semantic ambiguity, which is generally known as WSD, has proved to be more difficult than syntactic disambiguation.

Motivation

In our Research we attempt to develop a multilingual dictionary for all Indic languages plus English in an effective way, economizing on time and effort. We first discuss the disadvantages of language pair wise conventional dictionaries. In a typical bilingual dictionary, a word of *L1* is taken to be a lexical entry and for each of its senses the corresponding words in *L2* are given. It is possible that one sense of *Wi* in *L1* is exactly the same as one of the senses of *Wj* in *L1*. This means that *Wi* and *Wj* are

*Department of Computer Science, Assam University, Silchar.
** Department of Computer Science, Assam University, Silchar.

synonymous for a given sense. An example of this is dark and evil where one of the senses of dark and evil overlaps as for example in dark deeds and evil deeds. This phenomenon is abundant in any natural language. In a conventional dictionary, there is no mechanism to relate *Wi* with *Wj* in *L1*, though they conceptually express the same meaning. In turn, the corresponding words for *Wi* and *Wj* in *L2* are no way related to each other though conceptually they are. That is a major drawback, because of which conventional pair wise dictionaries cannot be used effectively in natural language application, especially when multiple languages are involved. The other disadvantage of the conventional dictionary is the duplication of manual labour. If an MT system is to be developed involving *n* languages, *n(n-1)/2 l*anguage pair wise dictionaries have to be created. For instance, if we consider *6* languages, *30* bilingual dictionaries have to be constructed. Finally, the effort of incorporating semantic features in *O(n2)* dictionaries is duplicated by *n/2* lexicographers-a wastage of manual labour and time

Indo WordNet

WordNets creation for languages other than Hindi is going on using the Expansion Approach. Figure 1 below shows the big picture of the Indo WordNet.

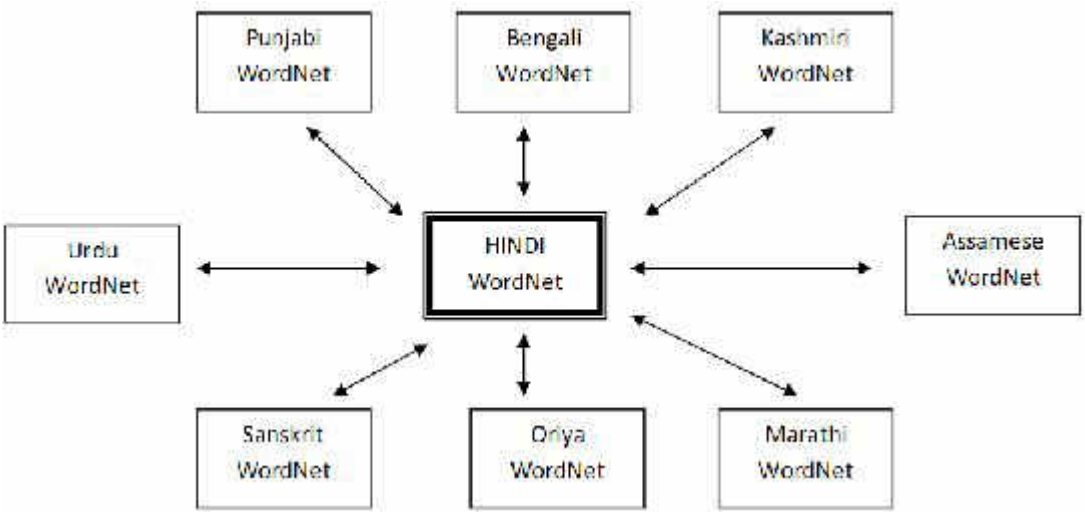

Figure 1: Linked Indo WordNet structure

Hindi WordNet Database Design

S. No.	Field Name	Purpose
01.	synset_id	Primary key: Uniquely identifies a concept/synset in the language
02.	concept_definition	The gloss / concept definition in a synset
03.	category_id	Foreign key from category table. Specifying if the concept is a noun, verb, adjective or adverb
04.	source_id	Foreign key from source table. Specifies the source from where the concept is taken
05.	Synset	inputting a set of synonyms representing that particular concept
06.	Example	examples of the given concept

Schema of the Dictionary Database

The Schema given below represents the Language L1 (e.g. Kashmiri). The others language L2 (e.g. Urdu) can be added by using the same Schema. The Language L1 and Language L2 can be linked by using the Dictionary Id.

Name	Type
Dic_id	Integer
Sense_id	Integer
Head Word	String
Lex_cat	String
Example	String

Fig 2. Schema of Language L1

In proposed multilingual dictionary we are using Hindi as a PIVOT Language. The others language can be added by using the linking method.

Conclusion

The work is in preliminary stage and needs certain improvements. So far we have developed a trilingual dictionary for Hindi, Kashmiri and English Language. We proposed a new pivot language-based method to create multilingual dictionary that can be used as translation resource for machine translation. Opposed to conventional methods that use dictionaries, our method uses WordNet as main resource of the intermediate language to select the suitable translation pairs. As a result, we eliminated most of the weaknesses caused by the structural differences of dictionaries, while profiting from the semantic relations provided by WordNet. We believe that because of the robust nature of our method it can be re-implemented with most language pairs.

References

Bar-Hillel and Yehoshua. (1964). *Language and Information*. New York:

Addison-Wesley, Chakrabarti, D., Narayan, D., Pandey, P., & Bhattacharyya, P. (2002). An Experience in Building the Indo-WordNet- A WordNet for Hindi. *1st Global WordNet Conference, Mysore, India.*

Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger. *Proceedings of the 6th Conference on Applied Natural Language Processing*, 2000: 224-231.

Brill, E. (1992). A Simple Rule-based Part of Speech Tagger. *Proceedings of the 3rd Conference on Applied Natural Language Processing*, 1992:152-155.

Cruse D.A. (1986). *Lexical Semantics,* Cambridge University Press.

Kulkarni, M., Dangarikar, C., Kulkarni, I., Nanda, A., & Bhattacharyya, P. (2010). Introducing Sanskrit Wordnet. *Global Wordnet Conference (GWC10), Mumbai, India.*

Fellbaum, C. (1998). *WordNet: An electronic lexical database.* MIT press.

Sinha, Manish, Reddy, M. & Bhattacharyya, P. (2006). An Approach towards Construction and Application of Multilingual Indo-WordNet. *3rd Global Wordnet Conference (GWC 06)*, Jeju Island, Korea, January.

Vossen, P. (ed.) (1999). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks for European languages*. Kluwer Academic Publishers, Dordrecht.