DRAVIDIAN WORDNET

(திராவிட சொல்வலையாக்கம்)

Dr.S.Rajendran
Professor & Head
Department of Linguistics,
Tamil University, Thanjavur 613010,
Tamilnadu, India
raj_ushush@yahoo.com

Abstract

Languages work as an integrated and interconnected system. Basically it connects the world with the human brain. The conceptualization of the world by perception and naming results in building the basic units of language, the words. This becomes the input into the brain from which the output of communication is made. Human beings more or less look at the world with similar experience and make abstractions about the world by conceptualization so that it is possible for them to exchange their ideas without difficulty. Language works at different levels between sounds, words and sentences. As sounds do not have meaning on its own human beings depend on words as basic units of communication. We visualize, conceptualize and name things based on our conceptual percepts. While doing so we make classifications of the objects of the world depending on their physical and telic properties. The psychology behind the abstraction leads to typology which in turn turned into ontology. That is the reason the base of WordNet's is ontology.

WordNet is a nonlinear lexical structure based on semantic features of individual words. It allows linking one word to another in a more meaningful way than in a conventional dictionary or thesaurus, since every word shares one or more of semantic features with the other in a language.

The primary use of wordNet is making use of it as a multilingual information accessing system through lexical items. The WordNet by its nature turns to be an ideal lexical accessing system as it links concepts with another concept by multifarious meaning relations. The wordNet is a lexical data base which has many practical utilities. One of them is to use of wordNet for translation. WordNet not only links one concept with another concept through semantic relations, but also captures the contextual meaning variations of a particular word i.e. the polysemy of a word. The wordNet has amble scope to link meaning with a context there by capturing the different meanings of a word contextually.

Development of Dravidian WordNet is an on-going project activity shared by Tamil University at Thanjavur, Amrita Vishwa Vidyapeetham at Coimbatore, and Dravidian University at Kuppam. The main objectives of the project are:

1. Developing an extensive and high quality multilingual semantic lexical database for Dravidian languages

2. Developing language-independent set of semantic concepts linking the language networks together

3. Standardizing semantic classification of information for all Dravidian languages and providing resources for development of applications

1 International Status

WordNet was originally conceived and developed as a lexical database for English on the basis of psycholinguistic properties. The major lexical categories like nouns, verbs, adjective and adverbs are organized in terms of sets of synonyms (synsets), each representing a lexical concept. A synset is a set of synonyms (word forms that have to the same or similar meaning) and two words are said to be synonymous if their mutual substitution does not alter the truth-value of a given sentence in which they occur, in a given context. For example, {car; auto; automobile; machine; motorcar} form a synset because they can be used to refer to the same concept. These synsets are interconnected by certain relations - lexical relations such as synonymy, antonymy, and semantic relations such as hyponymy (between specific and more general concepts) and meronymy (between parts and wholes).

The success of the English WordNet has paved way for the emergence of several projects with the aim of constructing WordNets in various languages and developing multilingual WordNets. EuroWordNet, a conglomeration of WordNets in European languages is an important project that has come up with a multilingual WordNet.

2 Dravidian WordNet

Dravidian WordNet is a natural chunk in the Indo-WordNet. It is only ideal that we should have Dravidian WordNet before we develop a larger Indo-WordNet, because the genetic relationship among the Dravidian languages can be maximally exploited in a more natural way. It allows, for example, a search tool to infer other terms, from the terms provided by the user and coming up with the most optimal search for retrieving information.

Among Indian languages, Dravidian languages such as Tamil, Telugu, Kannada and Malayalam share a number of lexicalized concepts in terms of morphology and semantics besides others as in typological and culture-specific features. Building a common WordNet for this family of languages makes it easier the task of developing an IndoWordNet.

The WordNet is a natural answer in machine translation systems. It has the potential to interpret source language words and come up with lexical equivalents in the target language in a more natural way as a bilingual does. This is particularly useful for users working in a second language who may not have appropriate knowledge of vocabulary. The network will also be used as a basic resource for supporting other applications. The semantic knowledge embodied in the network makes it suitable as a component in expert systems, question-answer systems, language learning systems and automatic summarizers. Dravidian WordNet with explicitly stated semantic features will become an essential lexical resource to be used in all practical NLP applications. It will give a boost for NLP research in India.

The resources produced by Dravidian WordNet will have a wide range of users who are interested in language learning, language generation, machine aided translation, language understanding, information retrieval, electronic publishing and the production of WordNets in other languages. The end users of the resources will be all those people who utilize the applications that incorporate Dravidian WordNet resources.

3 Design and Implementation

The design and implementaion of the Dravidian WordNet will be based on EuroWordNet as explained by Pike Vossen (1998). Designing and implementation of word net are the two major tasks assigned to computer scientists. To achieve them they have to work in collaboration with lexicographers and linguists. Once the lexicographers complete their work of collecting the data required for building the word net, the job will be handed over to computer scientists. The semantic information collected on lexical items from the basic building blocks

for the computer scientist to construct word net. As the words and meanings are related to one another and mapped as such in word net, it is but natural that the word net gives the impression of an on-line thesaurus. The word net automatically inherits the all the powers of a thesaurus. It also resembles an on-line dictionary as it provides meanings for lexical items. Being superior to these two tools, word net provides much more information that has been loaded in an on-line thesaurus as well as in an on-line dictionary.

4 Design of Individual WordNets

The task of developing the on-line database of a language can be conveniently divided into two interdependent tasks (Beckwith, Miller and Tengi, 1993). These tasks bear a vague similarity to the traditional tasks of writing and printing a dictionary:

1. To write the source files that contain the basic lexical data - the contents of those files are the lexical substance of WordNet.
2. To create a set of computer programs that would accept the source files and do all the work leading ultimately to the generation of a display for the user.

In line with English WordNet, the WordNet system can be divided into four parts based on the specific tasks assigned to them:

- Lexical resource system
- Compiler system
- Storage system
- Retrieval system

5 Expand/Merge approach

The WordNet database will be built (as much as possible) from available existing resources and databases with semantic information developed in various projects. This will be not only more cost-effective given the limited time and budget of the project, but also will make it possible to combine information from independently created WordNets. Two models will be involved in the built up.

Merge Model: the selection will be done in a local resource and the synsets and their language-internal relations will be first developed separately, after which the equivalence relations to Tamil WordNet will be generated.

Expand Model: the selection will be done in Tamil WordNet and the Tamil WordNet synsets will be translated (using bilingual dictionaries) into equivalent synsets in the other language. The WordNet relations will be later on adopted across languages.

The Merge Model will result in a WordNet that will be independent of Tamil WordNet, possibly maintaining the language-specific properties. The Expand model will result in a WordNet that is very close to Tamil WordNet but which is also biased by it. What approach should be followed also depends on the quality of the available resources.

After the first production phase the results will be converted to the Dravidian WordNet import format and loaded into the common database. At that point, various consistency checks will be carried out, both formally and conceptually. By using the specific options in the database, it is then possible to further inspect and compare the data, to restructure relations where necessary and to measure the overlap in the fragments developed at the separate sites.

6 Design of the multilingual database

The design of the Dravidian WordNet-database will be first of all based on the ontological structure of Tamil WordNet which in turn is based on a thesaurus for Tamil prepared by Rajendran (2001). Tamil wordNet relies on extensive preliminary investigations of the vocabulary of Tamil (Rajendran, 1976-2003) based on the componential analysis of meaning (Nida, 1975a & 1975b) and structural semantics (Lyons, 1977). Portions of this work have been compiled into a Tamil thesaurus (Rajendran, 2001). The Tamil thesaurus in electronic form represents the Ontological Structure of Tamil (shortly OST) vocabulary giving scope to take care of any kind of semantic/lexical relations that hold between lexical items.

The notion of a synset and the main semantic relations will be taken over in Dravidian WordNet. However, some specific changes will be made to the design of the database, which are mainly motivated by the following objectives:

1) to create a multilingual database;

2) to maintain language-specific relations in the WordNets;

3) to achieve maximal compatibility across the different resources;

4) to build the WordNets relatively independently (re)-using existing resources;

The most important difference of Dravidian WordNet with respect to a language specific WordNet is its multilinguality, which however also raises some fundamental questions with respect to the status of the monolingual information in the WordNets. In principle, multilinguality will be achieved by adding an equivalence relation for each synset in a language to the closest synset in Tamil WordNet. Synsets linked to the same Tamil WordNet synset will be supposed to be equivalent or close in meaning and can then be compared. However, we have to take into consideration the differences across the WordNets. If 'equivalent' words are related in different ways in the different resources, we have to make a decision about the legitimacy of these differences.

In Dravidian WordNet, we will take the position that it must be possible to reflect such differences in lexical semantic relations. The WordNets are seen as linguistic ontologies rather than ontologies for making inferences only. In an inference-based ontology it may be the case that a particular level or structuring is required to achieve a better control or performance, or a more compact and coherent structure. For this purpose it may be necessary to introduce artificial levels for concepts which are not lexicalized in a language or it may be necessary to neglect levels that are lexicalized but not relevant for the purpose of the ontology. A linguistic ontology, on the other hand, exactly reflects the lexicalization and the relations between the words in a language. It is a "WordNet" in the true sense of the word and therefore captures valuable information about conceptualizations that are lexicalized in a language: what is the available fund of words and expressions in a language. In addition to the theoretical motivation there is also a practical motivation for considering the WordNets as autonomous networks. To be more cost-effective, they will be derived (as far as possible) from existing resources, databases and tools. Each sites therefore will have different starting points for building their local WordNet, making it necessary to allow for a maximum of flexibility in producing the WordNets and structures.

7 The Database Modules

To be able to maintain the language-specific structures and to allow for the separate development of independent resources, we will make a distinction between the language-specific modules and a separate language-independent module. Each language module represents an autonomous and unique language-specific system of language-internal relations be-

tween synsets. Equivalence relations between the synsets in different languages and Tamil WordNet will be made explicit in the so-called Inter-Lingual-Index (ILI). Each synset in the monolingual WordNets will have at least one equivalence relation with a record in this ILI, either directly or indirectly via other related synsets. Language-specific synsets linked to the same ILI-record should thus be equivalent across the languages.

The ILI will be an unstructured list of meanings, mainly taken from Tamil WordNet, where each ILI-record consists of a synset, an Tamil gloss specifying the meaning and a reference to its source. The only purpose of the ILI is to mediate between the synsets of the language-specific WordNets. No relations are therefore maintained between the ILI-records as such. The development of a complete language-neutral ontology is considered to be too complex and time-consuming given the limitations of the project. As an unstructured list, there is no need to discuss changes or updates to the index from a many-to-many perspective. Note that it will nevertheless be possible to indirectly see a structuring of a set of ILI-records by viewing the language-internal relations of the language-specific concepts that are related to the set of ILI-records. Since Dravidian WordNet will be linked to the index in the same way as any of the other WordNets, it is still possible to recover the original internal organization of the synsets in terms of the semantic relations in WordNet.

Once the WordNets are properly linked to the ILI, the Tamil WordNet database will make it possible to compare WordNet fragments via ILI and to track down differences in lexicalization and in the language-internal relations. In this view, the ILI-records will be represented by a Tamil gloss. Below a synset-ILI pair, the language-internal relations can be expanded, as is done for the hyperonyms. The target of each relation will again be represented as a synset with the nearest ILI-equivalent (if present).

8 Conclusion

The theme of lexical semantics, computational lexicography, and computational semantics are altering rapidly. The availability of machine-readable resources and newly developed tools for analyzing and manipulating lexical entries make it possible to build a massive word net for a language. In present state of affairs it is quite feasible to build Dravidian WordNet. Building of a word net for Dravidian languages is an immediate requirement in the context of information technology equipped with internet in which the web sites in Dravidian languages are getting added up day by day.

References

Rajendran, S. 2001. *taRkaalat tamizc coRkaLanjciyam* [Thesaurus for Modern Tamil]. Thanjavur: Tamil University.

Rajendran, S. 2002. 'Preliminaries to the preparation of a Word Net for Tamil.' Language in India 2:1, www.languageinindia.com

Rajendran, S., S. Arulmozi, B. Kumara Shanmugam, S. Baskaran, and S. Thiagarajan. 2002. "Tamil WordNet." Proceedings of the First International Global WordNet Conference. Mysore: CIIL, 271-274.

Vossen P. (eds.) 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers**.**