Augmenting Indo-wordnet with Context

S. Rajendran
Tamil Universiy
Thanjavur
raj_ushush@yahoo.com

S. Arulmozi
Dravidian University
Kuppam
arulmozi@gmail.com

Abstract

It is difficult to interpret the meaning of a lexical item without context. Word net lists different senses of a word and provide definition and usage example for each sense. But like any sense enumerative lexicon it also does not provide any mechanism for the novel usage of a word. The polysemy found in verbs and adjectives convincingly tell us that we have to augment wordnet with context. Such mechanism will help us to condense senses listed under a word and allow us to interpret the senses of a word creatively or generatively.

1 Introduction

Word net as we understand is made up of synsets which are linked to each other by lexical and semantic relations in the background of ontology. Each synset represent a concept or a sense and the sense is given a description along with usage examples. For an end user word net serves both as a thesaurus as well as a dictionary. A user by typing a word in the interface slot can have a list of all the senses for the word. English word net for example lists 35 senses for the word *go* which includes 4 nominal senses, 30 verbal senses and one adjectival sense. Hindi on the other hand lists 2 nominal senses, 16 verbal senses and 2 adjectival senses for the word *chalnA* 'go'. Tamil lists 9 verbal senses for the word *poo* 'go'. There is no guarantee that only these are the possible senses for the word under consideration. As we know language is dynamic and not static. So there is always a possibility of expansion of the meaning of a word (i.e. addition of new senses) as the word may be used in new contexts. A static list of senses cannot capture this meaning expansion or generative use of words. The senses are also not compartmentalized; they are overlapping with one another. The lexicon which lists the senses of words can be called sense enumerative lexicon (SEL). SEL may not be able to capture the dynamic use of a word. It is in this respect argued in this paper that word net need to be complemented or augmented by a mechanism of condensing the senses listed under a word in the word net and providing a mechanism for interpreting novel senses in new contexts in which the word is being used.

Pike Vason (2001) points out the need for condensing meaning in word net. He states. "The matching of meanings across the word nets makes it necessary to account for polysemy in a generative way and to establish a notion of equivalence at a more global level." A context sensitive framework for lexical ontology like word net has been proposed by Velae and Hao (2007).

This paper is purely a theoretical one based on certain assumptions and there by point out or proposes a methodology to augment Indo-wordnet.

2 Limitations of Sense Enumerative Lexicon

Pustejovsky who argues for a generative frame work for a lexicon points out that lexical semantics should address the following issues (Pustejovsky, 1995:5):

(a) Explaining the polymorphic nature of language;
(b) Characterizing semanticality of natural language utterances;
(c) Capturing the creative use of words in novel contexts;
(d) Developing a richer, co-compositional semantic representation.

SELs are inadequate to account for the description of natural language semantics. Pustejovsky

points out that there are three basic arguments showing the inadequacies of SELs for the semantic description of language (Pustejovsky, 1995:39).

(1) THE CREATIVE USE OF WORDs: Words assume .new senses in novel contexts.
(2) THE PERMEABILITY OF WORD SENSES: Word senses are not atomic definitions but overlap and make reference to other senses of the word.
(3) THE EXPRESSION OF MUTLTIPLE SYNTACTIC FORMS: A single word sense can have multiple syntactic realizations.

Each of these consideration points to the inability of sense enumerative models to adequately express the nature of lexical knowledge and polysemy. Taken together, it would seem that the frameworks incorporating SELs are poor models of natural language semantics. A word may have contrastive or complementary senses. SEL lists contrastive senses as belonging to different words (i.e. as separate entries) and complementary senses as belonging to the same word (i.e. under same entry). Pustejovsky (1995: 38) restate the SEL's account of contrastive and complementary senses as follows:

> A Lexicon L is a Sense Enumeration Lexicon if and only if for every word w in L, having multiple senses $s_1,...,s_n$ associate with that word, then:
> (i) if $s_1,...,s_n$ are contrastive senses, the lexical entries expressing these senses are stored as $w_{s1},...w_{sn.}$
> (ii) if $s_1,...s_n$ are complementary senses, the lexical entry expressing these senses is stored as w{s1,...sn}.

Every ambiguity is either represented by (i) or (ii) above. Though Pustejovsky points out the advantage of this model of lexical description, he also states that the SEL model is inadequate for the purpose of linguistic theory.

3 Problem of polysemy in Verbal semantics

As we are making uses of limited number of verbs to express innumerable number of events and actions, verbs become significantly polysemous. So we are going to take up verbal polysemy to start with. As Tamil word net is only at it's of infant stage, we are going to make use of a representative SEL for Tamil (i.e *kriyaavin taR-kaalat tamiz akaraathi* (KTTA) (Dictionary of Contemporary Tamil) to serve our purpose. We will also make use of Generative Lexicon for Tamil (in manuscript form) written by Rajendran under a UGC sponsored project (Rajendran, 2010).

If we look at KTTA, we will find out that the number of senses enumerated under a verb vary from three to thirty approximately. Some verbs like *aTi* 'beat' and *pooTu* 'drop' acquire an enormous list of senses as they can collocate with a number of nouns forming different verbal senses. The different senses interpreted for them is based on the object noun with which they collocate with (for example *kaapi aTi* 'copy', *accu aTi* 'print' etc.). Such verbs behave like light verbs. So they show various senses based on the object-noun with which they collocate with. If we go through the dictionary and analyse the different senses listed under each verb, we will come to know that the context represented by the arguments of the verb (such as subject or agent-argument, object or patient-argument, indirect-object-argument, instrument-argument, location-argument) play a vital role in the interpretation of different senses for the concerned verbs. For the sake of illustration and to discuss the issue at hand a less polysemous verb *uTai* 'break' has been taken as an example. The different senses denoted by the verb are listed in the following table along with the sense descriptions and usage examples. The usage examples are analyzed for argument structure of the concerned verb.

Sr. no	Sense	Usage examples	Arguments	
			Subject	Object
1	துண்டாதல், பிளத்தல் ‘break; split’	1.திருடன் பூட்டை உடைத்து உள்ளே நுழைந்திருக்கிறான். ‘The has thief entered the house by breaking open the lock’	திருடன்	பூட்டு
2	(கட்டப்பட்டிருப்பதை அல்லது ஒட்டப்படிருபதை) பிரித்தல், மூடியாகப் பொருத்தப்படிருப்பதை திறத்தல் ‘break open (a bundle by snapping the string tied around), open (an envelope, a bottle, etc.)	1.துணிக்கட்டை உடைத்துப் ஒவ்வொன்றாக வெளியே எடுத்து விலை போட்டார். ‘He opened the cloth bundle, took out the cloth and wrote the prices 2.தபாலில் வந்த கடித்ததை உடைத்துப் படித்தார். He opened the envelope of the letter and read it. 3.அவருக்குச் சோடா உடைத்துக் கொடு ‘Open the soda bottle and give him’	1.அவர் 2.அவர் ‘he’ 3.நீ ‘you’	1.துணிகட்டு 2.கடிதம் 3.சோடா
3	(கட்சி நிறுவனம் போன்றவற்றைப் பிளத்தல், பிரிவுபடுத்துதல் Split; break up (a par-	கூட்டுறவு சங்கத்தை உடைக்க அவரி செய்த முயற்சிகள் வீணாயின. ‘The efforts he had taken to break up the so-	அவர் ‘he’	கூட்டுறவு சங்கம் ‘society’

	ty, an organization, etc)	ciety failed'		
4	(ரகசியத்தை, உண்மையை) வெளியாக்குதல் Make public (a secret, hidden facts, etc.); disclose.	1.அவர் யாருக்கும் தெரியாமல் வைத்திருந்த விஷயத்தை இப்படி உடைக்கலாமா? 'How can you disclose the secret he has kept to himself?' 2.உண்மையை உடைத்துச் சொல்லிவிட வேண்டியதுதான் 'I have to disclose the secret.'	1.நீ 'you' 2.நான் 'I'	விஷயம் 'matter' உண்மை 'truth'
5.	(கோடாலியால் மரத்துண்டுகளைப்) பிளத்தல் Split (logs)	விறகு உடைக்க ஆள் வரவில்லை 'The person to split the log has not come so for.'	ஆள் 'person'	விறகு 'log'

Table 1. Sense

The table reveals the fact that the object-noun of the verb determines the different senses assigned to the verb.

4 Problem of meaning interpretation of compound verbs

There are compound verbs in Tamil which are formed from a base by the addition of a verb which function as the verbalizer or whose function is to verbalize the base. The bases are generally nouns. Even a verb can be compounded with a verbalizing verb to form another verb. There are a number of verbs which are used to form verbs from nouns. Not all nouns can be added to a verbalizer and conversely not all verbalizers can be added to a noun; only a closed set of nouns can be collocated with a particular verbalizer. The compounds could be overlapping in their meaning as same nouns can be collocated with overlapping group of verbs. This leads to synonymy among compound verbs. Though the formation of verbs from N + V combination is a productive process, the nouns involved in the formation of compound verbs with reference to a particular verbalizer appear to be a closed set rather than an open set. But it is possible to recruit new members to a closed set which makes the process productive. Because of the closed nature of the nouns participating in the compound formation which results in the idiosyncratic nature of the resultant meanings, there need to be the listing of the compounds in the dictionary as soon as the compounds come into vogue. Instead of talking in terms of sets of nouns it is possible to talk in terms of semantic area or domain to which the nouns belong.There are thirty nine verbs in Tamil which can be claimed to function as verbalizers to form compound verbs from nominal bases.

Sl.No	Verbalizers with core meaning	Examples of Compound verbs in which the verbalizers form a part
1	*ati* 'beat'	*kan* 'eye' + *ati* > *kannati* 'wink'
2	*atai* 'get'	*mutivu* 'end' + *atai* > *mutivatai* 'come to an end'
3	*ali* 'give'	*paricu* 'prize' + *ali* > *paricali* 'award'
4	*aku* 'become'	*veli* 'outside + *aku* > *veliyaku* 'come out'
5	*akku* 'produce'	*coru* 'cooked rice' + *akku* > *corakku* 'cook rice'
6	*atu* 'move'	*kuttu* 'drama' + *atu* > *kuttatu* 'act'
7	*attu* 'swing'	*cir* 'orderliness' + *attu* > *cirattu* 'tend lovingly'
8	*arru* 'perform'	*pani* 'work' + *arru* > *paniyarru* 'work'
9	*itu* 'put'	*parvai* 'look' + *itu* > *parvaiyitu* 'inspect'
10	*uru* 'obtain'	*kelvi* 'hearsay' + *uru* > *kelviyuru* 'get to know'
11	*uruttu* 'trouble'	*tunpam* 'suffering' + *uruttu* > *tunpuruttu* 'cause suffering'
12	*uttu* 'give'	*ninaivu* 'rememberance' + *uttu* > *ninaivuttu* 'remind'
13	*etu* 'take'	*oyvu* 'rest' + *etu* > *oyvetu* 'take rest'
14	*eytu* 'obtain'	*maranam* 'death' + *eytu* > *maranameytu* 'die'
15	*el* 'accept'	*patavi* 'position' + *el* > *pataviyel* 'take office'
16	*eru* 'rise'	*cutu* 'heat' + *eru* > *cuteru* 'become hot'
17	*erru* 'raise'	*veli* 'outside' + *erru* > *veliyerru* 'expel'
18	*kattu* 'tie'	*itu* 'compensation' + *kattu* > *itukattu* 'make up'
19	*kattu* 'show'	*acai* 'desire' + *kattu* 'show' > *acaikattu* 'lure; tempt'
20	*kuru* 'say'	*puram* 'back' + *kuru* > *purankuru* 'backbite'
21	*kotu* 'give'	*peeccu* 'conversation' + *kotu* > *peccukkotu* 'initiate a talk'
22	*kol* 'get'	*totarpu* 'contact' + *kol* > *totarpu kol* 'contact'
23	*cey* 'do'	*vicaranai* 'investigation' + *cey* > *vicaranai cey* 'investigate'
24	*col* 'say'	*kol* 'lie' + *col* > *kol col* 'tell tale'
25	*tattu* 'pat'	*mattam* 'substandard' +*tattu*>*mattam tattu* 'degrade'
26	*patu* 'experience'	*vetkam* 'shyness' + *patu* > *vetkappatu* 'feel shy'
27	*patuttu* 'cause to experience'	*tunpam* 'suffering' + *patuttu* >*tunpappatuttu* 'cause to suffer'
28	*pannu* 'do'	*yocanai* 'thinking' + *pannu* > *yocanai pannu* 'think'
29	*par* 'see'	*vevu* 'spying' + *par* > *vevupar* 'spy'
30	*piti* 'catch'	*atam* 'obstinacy' + *piti* > *atampiti* 'become obstinate'
31	*puri* 'do'	*manam* 'marriage' + *puri* > *manampuri* 'marry'
32	*peru* 'get'	*oyvu* 'rest' + *peru* > *oyvu peru* 'retire (from service)'
33	*po* 'go'	*coram* 'adultery' + *po* > *corampo* 'commit adultary'
34	*potu* 'drop'	*cattam* 'sound' + *potu* > *cattam poTu* 'shout'
35	*muuTTu* 'make'	*kopam* 'anger' + *muuttu* > *kopamuttu* 'cause anger'
36	*va* 'come'	*valam* 'right' + *va* > *valamva* 'go round'
37	*vanku* 'get'	*velai* + *vanku* > *velaivanku* 'extract work'
38	*vitu* 'leave'	*muccu* 'breath' + *vitu* > *muccuvitu* 'breathe'
39	*vai* 'keep'	*ataku* 'pledge' + *vai* > *atakuvai* 'pledge'

Table 2. Verbalizers

It has to be noted here that all the verbalizing verbs are native Tamil words. Not all the verbs listed above are actually used as verbalizers. The number of compound verbs formed from each verbalizer also varies.

As inferred from the table the vebalizers or the light verbs depend on the preceding noun for the interpretation of the compounded meaning. Some of compounds formed thus find their place in the Tamil dictionary. But most of them are not listed in the dictionary as the process of this formation is productive. The question raised here is how are we going to list these verbs in the word net? Here again we need a generative mechanism to capture the polysemy in the light verbs.

5 Problem of meaning interpretation of adjectives

Adjectives in general depend upon the nouns they attribute for the interpretation of their meaning. The following examples in Tamil will illustrate this issue:

paccai poy (green lie) 'extreme lie'
paccai irattam (green blood) 'raw blood'
paccai kaaykaRi (green vegetable) 'raw vegetable'
paccai arici (raw rice) 'raw rice'
paccai miLakaay 'green chilly'
paccai taNNiir (green water) 'water (in general as opposed to cold water and hot water)'

The list is verb long showing idiosyncrasy in their interpretation. Though *nalla* 'good' can attribute any noun, its interpretation depends on the noun which follows it.

nalla peenaa 'good pen'
nalla peN 'good woman'
nalla katti 'good knife/sharp knife'
nalla aaciriyar 'good teacher/efficient teacher'

Here again we need a generative mechanism for the interpretation of adjectives.

6 A Proposal for Augmenting Indo-wordnet with Context

We are not going to adopt Pustovsky's model of generative lexicon (Pustejovsky, 1995) for our purpose. Adopting the methodology dealt by Pustejovsky to account for the polysemous structure found in the word net is difficult. We are planning for a different strategy that will suit word net and there by the contexts responsible for different senses of a particular word can be represented. If we again look at the table discussed above, we may infer that the semantic features of the nominal object determine the senses to be enumerated. It may be inferred that a set of items belonging to a domain of objects gives one sense to the verb and another set of items of object another meaning and so on. If it is possible to link these domains in an ontological tree, we may be able to infer the difference in the nominal object and there by assign different senses to the concerned verb. The nearness in the ontological hierarchy (which again is difficult to measure) may give rise to overlapping of senses. The distance in the ontological hierarchy may tell us how much the senses are apart. This will again help us decide whether the members of particular group of senses are complementary to one another demanding a common entry in the lexicon or contrastive demanding separate entries. We require a fine-grained ontological tree to implement this idea. This methodology will reduce the subjectivity in grouping senses under one or more entries in a lexicon. The context provided by the ontological tree can be exploited for the interpretation or generation various senses for a particular word.

7 Conclusion

Human categorization is neither a binary nor a context-free process. Rather, the criteria that govern the use and recognition of certain concepts may be satisfied to different degree in different contexts. Much work remains to be done on the current framework with the aim of a more formal treatment of how our approach serves to augment WordNet (or wordnet like resources) with concept descriptions that can be used both to categorize in context and to reason about those categorizations. WordNet is itself a little more than a classification hierarchy, and the conceptual functions we assign to its lexical entries serve much the same purpose (i.e. categorization and introspective reasoning).

References

Bouquet, P., Giunchigalia, F., Van Harmelen, F., Serafini, L. and Stuckenschmidt. H. 2003. C-Owl: Contextualizing ontologies. In proceedings of the 2nd International Semantic Web Conference, LNCS vol. 2870: 164-176. Springer verlag.

Fellbaum, (ed.) 1998. WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA.

Giunchiglia, F. Contextual reasoning. Special issues on I Linguaggi e le Macchine XVI: 345:364.

Levin, B. 1993. English Verb classes and alternations. University of Chicago Press, Chicago.

Pustejovsky, J. The Generative Lexicon. MIT Press, Cambridge, MA.

Rajendran, S. 2000, *taRaalat tamiz coRkaLanjciyam* (Thesaurus for Contemporary Tamil), Tamil University, Thanjavur

Rajendran,S. 2000. Strategies in the Formation of Compound Verbs in Tamil. International Journal of Dravidian Linguistics vol 29:2, 107-126

Rajendran, S.2003. "Creating Generative Lexicon from Dictionaries: Tamil Experience" In Rajeev Sangal el al (ed.) Recent Advances in Natural Language Processing.Myore: CIIL, 83-91.

Rajendran, S. 2010. Creating generative lexicon from MRDs in Tamil. (UGC project report), Tamil University, Thanjavur.

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

Subramanian, P.R. 1992. *Kriyaavin taRkaalat tamiz akaraati*. Cre-A, Chennai.

Vossen, Piek. 2001. "Condensed Meaning in Euro WordNet," in The Language for Word Meaning. Pierrette Bouillon, & Frederica Busa (ed.), Cambridge University Press, Cambridge.