Linguistically Annotated Dictionary (LAD) for Natural Language Processing – With special reference to Noun

Dr. R. Shanmugam,
Project engineer- GIST R&D, CDAC Pune.

ABSTRACT

Language is a social and linguistic phenomenon which grows continuously with all periodic changes. The grammar of Tholkappiar has provided considerable material for research in Tamil. At present in contemporary Tamil we come across many changes in both written and spoken usage and the progress of Language Technology have provided a new interface for researchers. One of the major areas in NLP is a Rule based approach which allows researchers to delve deep into the language however given the complexity of Tamil, one of the prime desiderata is a back end Dictionary that which plays a major role in reaching and disambiguating the output. Standard lexicons do not provide such information and only help us to conclude whether the word is valid or not. More complex information required for advanced areas like Word Sense disambiguation, Syntactic Parsing and Machine Translation such as meaning, antonym, etymology etc. is not accessible to the computer. Hence the need for a linguistically annotated dictionary (LAD) which will furnish more accurate information for NLP is must. The aim of this paper is to propose a design template for linguistically annotated Dictionary for Natural Language Processing with special reference to noun category.

Key words: Linguistically annotated dictionary (LAD), Natural Language Processing (NLP), Rule-based approach, Tamil-dictionary.

0. Introduction

Natural Language Processing (NLP) is an area which involves the creation of man-machine interfaces for the computer This can be done by concentrating on language Technology (the ways and methods to achieve NLP) and computational linguistics , the branch of linguistics which involves the techniques of understanding the linguistic properties which are scientifically interpretable. Rule based approach and statistical based approach are the two ways of achieving NLP. A rule based approach requires a number of generic rules to get the output whereas the statistical based requires certain rules, data and training to achieve the output. Human beings produce, understand and disambiguate the words and sentences only with the help of mental lexicon. This article is an attempt to make a near-by equivalent of mental lexicon which already exists in our brain and automatically getting updated through context or by the help of usage. This article aims at to produce a dictionary template for Linguistically Annotated Dictionary (LAD) which will work as a back-end dictionary for Rule-based Natural Language Processing (NLP) approach. The paper is divided into five sections including this Introduction. The second section discusses the semantic issues occurred in NLP analysis, the third section provides solutions to the issues by using LAD and the structure of LAD. The fourth part describes how LAD can be used for Parsers, WSD (word sense disambiguation), and Machine Translation systems and the final section will cover the conclusion which includes the ways and methods to build the LAD.

1. Semantic issues in NLP analysis

NLP needs a wide research to execute it and it starts with the fundamental levels of linguistics such as Phonology, Morphology, Syntax and Semantics. LAD is not required to handle the nominal analysis and issues on Morphology in which the simple root dictionary will be used to reach the output but LAD is much more essential to handle the influence of syntax and semantics on Morphology and the advanced analysis such as WSD, Machine Translation, Sandhi checker and Grammar Checker. For example the combination *paḻam* + *kutai* can generate two variants such as *paḻakkutai* fruit basket, *paḻaṅkutai* old basket. A Morphological Generator fails to handle this kind of combination since it requires additional information or rules to handle the issue. This is due to the meaning ambiguity with the word "*paḻam*" which has two different meanings "fruit" and "old". The synonymy of 'kaṉṉippeṇ' gives multiple meaning such as unmarried young women, spinster, virgin and this can be determined only by the **(**linguistic or extra-linguistic**)** context.

2. Solutions

The root word dictionary cannot provide more information about the word except that it belongs to a separate category and valid but this information is not enough (sufficient!!) to develop word sense disambiguation and Machine Translation systems. Extra information is required to solve the ambiguities and get the actual meaning of the input. This information has to be stored in our dictionary so that the execution of word sandhi can be done with the help of LAD. For the first example *paḻam+* kūṭai the word *paḻam* has one more meaning that which stands behind the second output. On the other hand, in the example two, the multiple senses of the word 'kaṉṉippeṇ' generates multiple outputs. So maintaining a semantic extension of each word is important to solve the semantic issue. A Morphological generator will have no trouble to generate the multiple outputs for ambiguous pairs provided it is enabled with LAD. Semantic features such as connotation, denotation and domain play a vital role in disambiguation. LAD provides all the semantic extensions of each word and hence retrieving the same is not an issue for the application.

2.1. Linguistically annotated Dictionary (LAD)

A "word' is not only a word, but it has many things inside starting from etymology. Designing a dictionary for word reference is different from doing the same for Natural Language Processing. As discussed earlier, the root word dictionary will not be sufficient to handle ambiguity or advanced NLP analysis. LAD is a dictionary with grammatical and semantic information which will cover all the meaning of a word including its denotative, connotative and domain wise. It will work as a back end dictionary to NLP applications. This will help us to increase the accuracy and quality of our NLP applications. The present article will only talk about the LAD of Tamil Noun but the same needs to be designed for all the categories which will have semantic value, especially verbs, the backbone of sentence formation. This will be useful to enable the NLP engine for handling lexical and structural ambiguities.

2.2. Structure of LAD

The Structure of LAD decides or determines the accuracy of the Language Tool. Each Head word should contain a semantic information such as denotative, connotative, synonyms, antonyms, idioms, proverbs, domain (domains will be used to cover a domain wise meaning of each noun) and syntactic information such as Collocation, parts of speech, sub-categorization such as abstract or common, count or mass noun, animate or inanimate and pronunciation with different types of mood which will be useful to capture the different meaning variants, IPA which are helpful to cover the prosodic features for Text to speech. The category wise meaning will also be listed in the LAD. Since nouns have been taken for the present article the same have only been focused here. So building a LAD is equal to building a word structure with syntactic and semantic information which is always absent in the root word dictionary. XML is one of the preferable formats for LAD. Building a dictionary with all semantic information is not achievable hence this article only aims **at** to cover some semantic information because syntactic and semantic information of each word is one of the major strength of human's mental lexicon and it plays a massive role in understanding the meaning and disambiguating the sentence. The following table shows the structure of LAD.

	Word 1	Word 2	word3
Root	vīīṭu(home)	rōja(rose)	cuṉāmi (Tsunami)
PoS	peyar	peyar	peyar
Etymology	-	-	-
sub-category	Common-concrete-count-inanimate-neuter	proper-concrete-count-inanimate-feminine	Common-concrete-count-inanimate-neuter
IPA& Pronunciation	ʋi□ɖʊ	ɾ o□dʒa:	tʃʊna:mɪ
denotative	vāl̲umiṭam	malar	āl̲i pēralaikaḷ
connotative	kuṭumpam	kaṭal ciṉṉam	al̲ivu,sōkam
Domain1	-	-	-
Domain 2	-	-	-
Idioms and proverb	vīṭu pōla eṇṇutal	rōjā pōṉṟu meṇmai	perum palattuṭaṉ kūṭi
synonyms	illam		āl̲i pēralai
Antonyms	-	-	-
collocation	-	civappu	-
Compound stem	vīṭṭu	rōja	cuṉāmi

Note: the sub-category has been given for the primary meaning and it has to be updated and maintained for each meaning extension. Domains may change the POS of the particular entry that information have also to be maintained. This above table is only a design template of LAD; if required**,** it has to be updated with more categories.

2.2.1 Denotative & connotative

Denotation means the referential meaning(s) of the word i.e. the explicit definition of the word as listed in a dictionary. For example the denotative meaning of the word **cuṉāmi** is a series of water waves caused by the displacement of a large volume of a body of water, generally an ocean or a large lake. On the other hand the same word will have a connotative meaning of danger and destruction. This connotes the social overtones of "**cuṉāmi**".

2.2.2. Domain

A word is not restricted to its connotative and denotative meanings. It very often acquires a tertiary set of meanings in terms of the domains or areas in which it is used e.g. astronomy, art, literature etc. For example the word "*cīṭṭu*" will have multiple meaning depending on the domain. It could mean receipt in business or economic domain. The same would mean as ticket in transport domain e.g. *payaṇac cīṭṭu* (Travel ticket) and it also playing cards in the ludic domain. Thus maintaining the domain wise meaning is very helpful to handle the meaning extension and provide the full semantic profile of the word.

2.2.3 Collocations: Idioms and Proverbs

Idioms and Proverbs are fixed expressions and are major issues in NLP applications since the actual meaning of the word changes and modified within the collocation. This is especially difficult in the area of translation. MT systems based on Direct Translation method fail to handle idioms and proverbs. In many cases the connotative meaning of the word will be considered to get the actual meaning but in root word dictionary there is no option to get connotative meaning of each word. Thus the proverb *miṉṉuvatellām poṉṉalla* (all the glitters is not gold) meaning that not everything that looks precious or true turns out not to be so. It is applicable for human, places and things etc. The original context of this proverb decides the subject. Hence MT system needs extra world knowledge to handle them.

3. How LAD can be used for NLP

Creativity is one of the important features of Language which constantly changes with use. This is the major stumbling block for NLP. A disambiguator is a much needed tool which will try to bring the original meaning of the sentence especially for Machine Translation. Lexical ambiguities can be handled by using some rules but to solve structural ambiguity a high quality syntactic parser is a must and to develop a good syntactic Parser there is a need for an accurate Morphological parser which in turn can handle lexical ambiguities For example the sentence "*avarkaḷōṭu iṇaiya mānaṭu ceṉṟāṉ*" gives two meanings that "he went to conference to join with them" or "he went to Internet conference" it happens due to the ambiguous word *iṇaiya* which acts as a part of the noun *iṇaiyam* and the infinitive form of the verb *iṇai*. A naive Morphological Parser may fail to guess the nature of this particular entry and it requires additional rules to handle such inputs .Generally parsers do not have much knowledge of collocation needed to capture the valid combination thereby making it impossible to generate a single meaning for this sentence without referring to the context. However the LAD provides the option to have a second meaning of this example by mentioning *mānāṭu* as a collocation of the word *iṇaiyam*. Since this is a well-known pair of term it can be handled efficiently and by doing so resolve the contextual analysis of both outputs and the system will procure information as to the exact meaning of the sentence.

In some places the word's actual form can be used to use the actual nature of that particular word for other word. For example the word "*praccāra pīraṅki*" means a person who is sound in canvassing and audible to many people, so the actual nature of the word *pīraṅki* has to be considered that which shows the power of the person who canvasses more effectively than others. Machine Translation systems fail to handle such kind of combinations because of the lack of additional information for each word. Having synonyms and antonyms is much more useful to develop an accurate NLP engine because in some cases antonyms will be used to get the actual negative meaning of the utterance in different pitch variation occurred due to tone of the speaker.

4. Conclusion

LAD will serve as a finer prerequisite for advanced NLP research and building the same takes more time but it will support as a strong base for any NLP application either it may be a Text or Speech Technology product. All the fields mentioned can be stored in the shape of an XML database which will allow the user to extract information efficiently. A prototype of such a dictionary has been commenced by CDAC GIST. Since the effort required to build such a lexicon is huge, as is the practice, a crowd sourcing model will be deployed to build lexicons for Indian languages. The database so created will allow linguists to create support for their own language and after the proper validation is done by the validators and administrators the data will be uploaded and shared. The project, at present is in alpha stage. LAD will cover the syntactic and semantic information of each word so it will be helpful to increase the accuracy level of NLP products far better than before.

References

1. CREA – *Modern Tamil dictionary for Tamil-Tamil-English* – by CREA 2009.

2. Chomsky, Noam. *Logical syntax and Semantics – Their linguistic Relevance*- Language Volume 31. No.1 (Jan-Mar 1995).

3. Cook, Vivian – Mark Newson. *Chomsky's Universal grammar*- Blackwell Publications

4.Kothandaraman-.Pon. *A grammar of contemporary literary Tamil.* IITS -1997.

5 Kothandaraman-.Pon. *Tamil studies selected papers.* Ambuli Publication-2001.

6. Krishnamoorthy, Badriraju. *The Dravidian Languages.* Cambridge University Press -2003

7. Moravcsik JME. *Sub-categorization and abstract terms*-in Foundations of Language Vol.6 No.4 1970.

8. Verma S.K, N. Krishanaswami. *Modern Linguistics -An Introduction*- Oxford University Press – 1989

9. en.wikipedia.org/wiki