**SYNSET CATEGORIZATION FOR SYNSET CREATION:**

**A PROPOSAL**

**Aadil Amin Kak Nazima Mehdi Aadil A. Lawaye Mansoor Farooq Farooq A. Shiekh**

Dept. of Linguistics

University of Kashmir

{aadilaminkak,nazimamehdi,aadillawaye}@yahoo.com

{khmansoor003, farooqahmad84}@gmail.com

Abstract

The present paper underlines some problems faced during synset ranking and categorization. Taking a view point of resolving those issues by categorization, the paper proceeds to postulate a tentative categorization system of synset categorization. The system is explained and how the problems can be resolved is dwelt upon. Furthermore, the paper also tries to constrain the concept of ‘natural lexeme’ and certain other related concepts.

1. Introduction

WordNet is a lexical database for different languages which group’s words into sets of synonyms called synsets, provides short general definitions and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications.

Indradhanush WordNet, a part of IndoWordNet has the basic purpose of developing a concept-based multilingual dictionary for Indian languages. It involves the three following steps

-concept development

-providing example/examples of the given concept, and

-inputting a set of synonyms representing that particular concept.



In the above given snapshot all the end-products of the three above mentioned processes are observed.

Going back into history, Hindi wordnet was the first and most comprehensive wordnet in Indian languages. So, when the concept of an indowordnet was conceived it was proposed to take Hindi as the pivot language. However, on observing differences among Indian languages, it was proposed that core synsets and synsets common to all Indian languages (languages which were being worked on) would be identified and worked on in the beginning. After carrying out the initial work it was finally put forth that taking the Hindi wordnet as the base (the pivot), the first step should be to see if that concept (the Hindi concept) is a natural lexeme in the individual language or not. In other words, all the language groups are ranking Hindi synsets and observing whether for a particular given concept is there a natural lexeme present in their Target language (Kashmiri, Marathi, Bengali, Panjabi) or not. This would in a way be the first step of getting a set of concepts which are common to all the languages being worked on. So, presently every language group is doing synset ranking by understanding whether for a particular given (Hindi) concept is there a natural lexeme present in their respective languages or not.

1. Problems

The problems faced in this first step of generating a set of common concepts are the following.

* 1. Proper Nouns

One of the problems is ranking Proper Nouns. For example, what should be the criteria of inclusion of a country in the Kashmiri wordnet? If we consider China, Somalia, Iraq, Burma, Bhutan, Togo, Croatia etc, which countries should be included as natural lexeme of Kashmiri and what should be the criteria for inclusion? Should we take China as a natural lexeme because Kashmir has had historical trade links with China from the Silk Route perspective? Should we take Somalia because of the famine and war which brought it into news? Incidentally, how many people knew there was a place Mogadishu before the famine and war. And what about Togo, Croatia, etc? Coming down to states, we have different states within countries even in India, and within states we have different districts, and within districts we have different villages and then galis and kuchas (lanes and bylanes), what about them?

* 1. Borrowing

Another problem is borrowing. Linguistically speaking, borrowing is a very fluid notion. What is borrowed for some may not be borrowed for others. This is mostly observed in technical jargon. Borrowing is also time dependent, what was not borrowed earlier, may be borrowed now, and what is borrowed now may no longer exist later. For example computer, pant, aasmaan (sky). Another important point is that borrowing may also involve a change in concept, which adds to the problem. For example, in Kashmir the word ‘engine’ from English is used very commonly for ‘Road rollers’, the word ‘motor’ for the older generation is any vehicle, the word ‘Punjabi’ is used for anyone who is from any part of India except South and North-East except for labourers who are called ‘Biharis’.

One more problem is that of literacy. We look at literacy as an issue here when we take literacy to mean being ‘literate’ in languages other than the mother tongue. This also plays a role in conceptualization. The verbal repertoire of the literate class will be somewhat different which also depends on the attitudes, educational level and certain other social and psychological features. This will definitely have an impact on synset building, whether for better or worse, is a different question, but we can’t ignore the fact that the effect will be there. Words like cockpit, level, understanding, etc are some examples.

* 1. Hindi as pivot

Another valid point which has to be considered here is that Hindi being taken as a pivot will also have its own problems. This will invariably cause the person working on other languages be influenced, however hard he/she tries otherwise. The problem will invariably be looking at the concept from a ‘hindi’ position which may influence his decision of inclusion of the concept in his language. The degree of ‘seeming-to-be-equivalent’ between Hindi and the target language can also be influenced by the Hindi proficiency of the lexicographer and his/her attitude towards Hindi and the language he/she is working on.

* 1. World view

Another point which is again to be taken into consideration is the world view of the lexicographer which can be influenced by a number of factors, chief among them is religion. Now, to what extent can we carry on the ‘Natural Lexeme-basis-of-finding-the-common-concept Theory’ in this matter? Even within religion, proficiency in religion also plays an important role. Now, considering or ignoring deities, prophets, gurus, saints as natural lexemes in ones language can be a very daunting task, especially considering the sensitive nature of the people in the Indian subcontinent in those matters. One has no other option but to look for other options.

* 1. Conceptual inadequacies

Furthermore, conceptual inadequacies are another point which should be taken into consideration. ‘eik prakar ka something….’seems to be very non-professional and vague. Another thing which also crops up is that when conceptual inadequacies arise then it becomes difficult to identify what the concept is all about, sometimes even with the given Hindi examples, especially for those who live in the non-Hindi belt.

Overall, after carefully going over the problems it seems that most of the problems hinge around the concept of Natural lexeme. So, the first and foremost step before arriving at the common concepts theme is that the notion of natural lexeme should be clarified. It should be understood, as explained above that the understanding and defining a natural lexeme is not easy. The concept of natural lexeme, as far as we understood it, is not homogenous and varies across time, space, belief system, etc. The second step after this should be a strict and formal categorization which should be followed. This will undoubtedly help in the present future processes of indowordnet creation.

1. Our Proposal

In order to solve such problems we are attempting to propose a proper and formal categorization, and within this categorization the concept of natural lexeme should also be made clear.

Overall, the synsets in question can be divided into the following categories.

General Lexicon

Natural Lexeme

Borrowed Lexemes

Religious Concepts

Transliterated/Coined Terms

Common Natural Lexeme

Region Specific Natural Lexeme

Commonly Borrowed

Specifically Borrowed

Culture Specific Natural Lexeme

Geographic Terms

Place Names

Flora and Fauna

* 1. Total Lexicon

This will include all the concepts of the language including the transliterated/coined concepts which are imported from other languages. The combination of the total lexicon of all participant languages will be the total lexicon of Indowordnet/Indradhanush.

* + 1. Natural Lexeme

Coming to the natural lexeme concept, it is proposed that all concepts which native speakers feel

a. are native

b. identify with them and

c. relate to them

 be considered as Natural lexemes. It is important to point here that this concept should not be connected to frequency. For example, apple is a native lexeme for Kashmiris, while banana, mango, orange though very frequent cannot be considered native lexemes from a Kashmiri point of view. Consider Guitar, though many people might know about it, is not a natural lexeme for a Kashmiri, for whom Rabab would be a natural lexeme.

In order to resolve any ambiguity regarding if a concept is a natural lexeme of a borrowed one, a a natural lexeme should fulfill all the above mentioned criteria. Taking the example of kashmiri, if kashmiri native-speakers think that the concept is native and identify and relate to it then only is the concept a natural lexeme. Coming to more examples, consider the Chinar which is historically claimed to have been imported from Iran, Though the concept appears to be borrowed, it is a kashmiri natural lexeme because it fulfills the three above mentioned criteria. If we consider a car, the word and the concept is very common but it cannot be a natural lexeme of kashmiri. The concept of the natural lexeme should become more clear as we move ahead.

Natural Lexemes are divided into mainly 2 subtypes: Common Natural Lexemes and Region Specific Natural Lexemes

* + - 1. Common Natural Lexemes

These include the most common natural lexemes around us. These lexemes are presumably present in other languages as well. E.g. the concept of sleep, eat, drink, soil, sky, cow, etc.

* + - 1. Region Specific Natural Lexemes

These are natural lexemes specific to the region/regions of the language in question. These are further classified into 4 major subheadings: culture specific natural lexemes, geographical natural lexemes, Place Names and Flora and fauna.

* + - * 1. Culture Specific Natural Lexemes

These are the natural lexemes which are culture specific and would take care of all cultural idiosyncrasies e.g. pheran, kangIr, samavar in Kashmiri.

* + - * 1. Geographical Natural Lexemes

This subheading was considered considering different geographical conditions in India and cater to geographical and geological peculiarities of different places.

* + - * 1. Place Names

The lexemes which are the names of the places under the influence of the language but the lexemes are natural to that language, e.g. Srinagar, Hazratbal, Sopore etc. in Kashmiri.

* + - * 1. Flora & Fauna

The lexemes which are the names of plants and animals specific to the area of influence of the language but the lexemes are natural to the language, e.g. hangul, soi, etc. in Kashmiri.

* + 1. Borrowed Lexemes

These are the lexemes which are not natural to the language concerned but are borrowed from other languages. Borrowed lexemes are further subdivided into two types viz

* + - 1. Commonly borrowed lexemes

These are the lexemes which are commonly borrowed by all the languages. E.g. banana, computer, Television, car etc.

* + - 1. Specific borrowed

These are the lexemes which are specifically borrowed by a particular language and are usually parts of technical jargon. E.g. wiper, over-haul will be borrowed lexemes for mechanics, RAM, Hard Disk, Processor for computer workers, etc.

* + 1. Theological Concepts

In this group will be included all concepts relating to religions, sects, cults, etc, all who have some religious tinge. Deities, gods, goddesses, religious places, etc will be included in this group irrespective of their being common natural lexemes of any individual language/languages.

* + 1. Transliterated/Coined Terms

This will include all the terms which are not natural lexemes, nor borrowed lexemes and nor theological lexemes. They, in a way will incorporate all foreign concepts other than those which don’t fall in the above mentioned groups.

One more important point which needs to be added pertains to the ambiguity of concept. It is proposed that apart from making unambiguous definitions of concepts some way should be evolved which would involve componential analysis as well. This would definitely make the whole process less ambiguous, more formal and constraint driven.

1. Conclusion

The main reasons for going for this categorization are the problems faced while working in the Indowordnet especially the Indradhanush part. Kashmir, as we feel is true for other regions of india as well, is culturally, geographically, ethnically distinct. To put together distinct cultures in a common framework is a daunting task, and this categorization should be the first step towards this. Though the proposal is tentative and very basic and needs more work and ironing out, this should guide us towards a very formal approach to the Indowordnet development. As observed in discussions in different indowordnet workshops, it is not easy to capture the diversity of India with its languages, dialects and variety, this proposal is a humble attempt at capturing some of the nuances. It is also felt that taking these categories into consideration would make it easier to rank synsets, and simultaneously make it easier to create broad-domain specific translation systems.

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.

Cruse D.A. (1986), *Lexical Semantics,* Cambridge University Press. Kulkarni M., Dangarikar C., Kulkarni I., Nanda A. And Bhattacharyya P. (2010), *Introducing Sanskrit Wordnet*, Global Wordnet Conference (GWC10), Mumbai, India.

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

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