Experiences in building the Nepali WordNet - insights and challenges

Alok Chakrabarty
Department of Computer Science
Assam University Silchar
mcscalok@gmail.com

Bipul Syam Purkayastha
Department of Computer Science
Assam University Silchar
bipul_sh@hotmail.com

Arindam Roy
Department of Computer Science
Assam University Silchar
arindam_roy74@rediffmail.com

Abstract

Machine translation in Nepali language is in an infant stage in comparison to other scheduled languages of India like Hindi, Sanskrit, Tamil, etc. One of the major reasons behind this is the non-availability of rich lexical resources in Nepali. The Nepali WordNet is thus an endeavour to prepare a rich lexical resource for the Nepali Language for effective machine translation and to facilitate the development of Information and Communication Technologies in Nepali. The endeavour is inspired by the famous English WordNet and the Hindi Word-Net. In the present paper we discuss some of the preliminaries involved in this attempt like the expansion approach of WordNet creation, the linguistic challenges involved, WordNet creation tool interface and the synsets' storage structure. We also discuss some of the special characteristics of the Nepali language.

1   Introduction

According to Miller, et al. (1993), "WordNet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory." In a WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct lexical concept or sense. These synsets are interlinked by means of conceptual-semantic and lexical relations. With each synset, WordNet provides a short and general definition for that sense. As it stores the lexical information in terms of word meanings whose organization conforms to the current psycholinguistic theories of human lexical memory it can be termed as a lexicon based on psycholinguistic principles.

The Nepali WordNet is an attempt to prepare such a lexical reference system for the Nepali language along the lines of the famous English WordNet (Fellbaum, 1998; Miller, 1995) and the Hindi WordNet (Chakrabarti et al., 2002) so that it can be used as a tool for enhancing the performance of Machine Translation and cross lingual information retrieval systems involving Nepali and to facilitate the development of various Information and Communication Technologies in Nepali.

The roadmap for the rest of the paper is as follows:

Section 2 is on some of the special characteristics of the Nepali language. Section 3 is on the expansion approach of WordNet creation, the *relation borrowing* concept and the challenges faced therein. Section 4 presents a discussion on the WordNet creation tool interface and the synsets' storage structure, and finally Section 5 concludes the paper.

2   Special Characteristics of Nepali

Nepali (नेपाली) is a language in the Indo-Aryan branch of the Indo-European language family with approximately 40 million speakers in Nepal, Bhutan, Myanmar and parts of India. It is the lingua-franca of Nepal and is one of 23 official languages of India, incorporated in the Indian constitution. It has official language status in the Indian states of Sikkim and in West Bengal's Darjeeling district. Further it is widely spoken in

the Indian states of Uttaranchal and Assam (Nepali language, 2009).

Unlike English, Nepali, like Hindi and its ancestor Sanskrit is a *Subject Object Verb* (SOV) language, i.e., in Nepali, the subject, object, and verb of a sentence usually appear in that order. For example:

Sentence:	उसले	मेरो	केरा	खायो।
Transliteration:	*usle*	*mero*	*keraa*	*khaayo.*
Gloss:	he	my	banana	ate.
Parts:	Subject	Object		Verb
Translation:		He ate my banana.		

Nepali is written left to right in the Devanagari script. It is written phonetically, that is, the sounds correspond almost exactly to the written letters. Nepali has many loanwords from Arabic and Persian languages, as well as some Hindi and English borrowings.

A deviating feature of Nepali among the Indo-Aryan languages is in terms of grammatical gender. Nepali possesses an "*attenuated gender*" system in which the gender accord is typically restricted to non-human female animates (Masica, 1991). For example:

[Human: Male, female]

Sentences:	केटो आयो,	केटी आई
Transliterations:	*keTo aayo,*	*keTi aaee*
Translations:	Boy came,	Girl came

[Non-human: Male, female]

Sentences:	गोरु आयो,	गाइ आयो
Transliterations:	*goru aayo,*	*gaai aayo*
Translations:	Bull came,	Cow came

The above issue raises problem in deciding some Nepali synsets as discussed in the next section. The old English (Anglo-Saxon) had such kind of distinction in grammatical gender but modern English is normally described as lacking grammatical gender.

As per verb morphology, Nepali has only two genders *masculine* and *feminine* for nouns. Gender in Nepali is a syntactic property. For both the genders a common pluralizing suffix 'हरु', '*haru*' can be used for nouns in Nepali, like केटाहरु, '*keTaaharu*' (boys), केटीहरु, '*keTeeharu*' (girls). Unlike English its usage is not mandatory and may be left unused if plurality is already indicated in some other way like by explicit numbering, or agreement (Cardona and Jain, 2003).

3   The Expansion Approach of WordNet creation

The Expansion Approach of WordNet creation is an effective method for creation of a new WordNet for a language from a well-established one. It was first proposed within the EuroWordNet project (Vossen, 2002). Thereafter it has been used by a number of WordNet development teams for the creation of new WordNets. Examples include the WordNets for Spanish, French (Vossen, 2002), Hungarian language (Alexin et al., 2006), Hindi, Marathi (Sinha et al., 2006), etc.

In the Expansion Approach, synsets of a pre-existing WordNet are understood by the lexicographer and the corresponding target language synsets expressing the same sense are created.

The Nepali WordNet also is under construction using this Expansion Approach as a consortium project with IIT Bombay. The WordNet is presently at an infant stage.

Because of the high degree of similarity between the two languages Hindi and Nepali, the Hindi WordNet has been used as the pivot WordNet in this approach. In the software tool in use for this purpose, almost all Hindi synsets are linked with the corresponding synsets of the English WordNet. The English WordNet is thus in use presently for the resolution of ambiguity of senses for the synsets in the Hindi WordNet (in a linked fashion with the Hindi WordNet). Once done with the Hindi synsets the process may further be extended to include additional synsets in Nepali WordNet from the English WordNet using English as pivot language.

Henceforth we will refer to the Nepali WordNet, Hindi WordNet and English WordNet as NWN, HWN and EWN respectively.

The main idea behind the Expansion Approach is the concept of *relation borrowing*. It refers to the relation establishment for one WordNet using the relations of another WordNet. The technique is automatic for semantic relations which link concepts or senses, but semi-automatic for lexical relations which link individual words. Different cases of relation borrowing from HWN to NWN are as follows:

a. *A sense is present in both Nepali and Hindi*:
Since Hindi and Nepali belong to the same linguistic family (Indo-Aryan) and exist in almost identical cultural setting, this is the commonest case. In this case relations are established in NWN for that sense.

b. *A sense is present in Hindi but not in Nepali*:

In this case the relations will not get established for this sense. For example, {आठवारा [*aaThwaaraa*, a period of eight days, like Monday to Monday]} is a sense in Hindi for which there is no corresponding sense in Nepali. Such a sense may be termed as a Hindi specific sense. In such case a lexicographer should adopt the following methods:

1. Transliteration
2. Use of multiword expression (short phrases)
3. Coining of new words

The steps should be used in the mentioned order of priority.

c. *A sense is present in Nepali but not in Hindi*:

Such a sense may be termed as a Nepali specific sense. The relations for such a sense in Nepali have to be established manually. For example, {पेवा [*pewaa*, a portion of the property of family owned by a female member]} is a sense in Nepali which does not have any correspondence in Hindi.

Similar cases of relation borrowing also exist from EWN to NWN.

WordNets deal with the content words, or open class category of words. Thus, the NWN also contains the open category of words viz. Noun, Verb, Adjective and Adverb. As per HWN, in the NWN also there are various semantic and lexical relations. In total there are 16 such relations (Hindi WordNet Documentation, 2009). They are:

1. General relations between synsets: *Hypernymy, Hyponymy, Meronymy, Holonymy*

2. General lexical relations: *Antonymy, Gradation*

3. Verb specific lexical relations: *Entailment, Troponymy, Causation*

4. Cross parts of speech linkage

a) Linkages between nominal and verbal concepts (Semantic relations): *Ability Link, Capability Link, Function Link*
b) Linkage between nominal and adjectival concepts: *Attribute (semantic relation), Modifies Noun*

c) Linkage between adverbial and verbal concepts: *Modifies Verb, Derived From*

3.1 Challenges faced in the Expansion Approach

We now discuss about the challenges that are faced in using the Expansion Approach. For each case of relation borrowing discussed above, following challenges were encountered:

Case a: Even if this is the commonest case and for most of the synsets it is easier to do so, a specific challenge arises for this case also:

1. When a sense in Hindi, expressed through a single word expression cannot be expressed so in Nepali and requires a multi-word synthetic expression. For example the sense of लड़ना, '*laRna*', (quarrel), a verb in Hindi, is expressed in Nepali by the synset {झगडा गर्नु [*jhagRa garnu*], भनाभन गर्नु [*bhanaabhan garnu*], कलह गर्नु [*kalah garnu*], कुटाकुट गर्नु [*kuTaakuT garnu*], कुटामारी गर्नु [*kuTaamaari garnu*]} all synonyms requiring two-word noun + verb expression [गर्नु, '*garnu*', is the verb]. In such cases the way-out is to put the multiword expressions in the synset of the target WordNet (here NWN) joined by underscores as under:

{झगडा_गर्नु, भनाभन_गर्नु, कलह_गर्नु, कुटाकुट _गर्नु, कुटामारी_गर्नु}

However problems may arise if the multiword expression becomes too long.

Case b: Out of the three solutions proposed above, transliteration is comparatively easy and straightforward. In many languages adoption of new words by transliteration has been done. Examples include the adoption of words like 'mobile', 'typewriter', 'cycle', 'station', 'coffee', 'machine', 'driver', 'pilot', 'table', 'chair', etc. from English to Hindi and Nepali and words like 'cheetah', 'brahmin', etc. from Hindi to English. For many of the words mentioned above, for one language there existed no corresponding word with exact sense in the other language, like 'mobile', 'typewriter', 'cycle', 'station', 'coffee', 'driver', 'pilot' of English had no counterparts in Hindi and Nepali. However because of the difference in pronunciation due to regional factor, many of these words got slightly changed or deformed. For example:

'doctor' = डाक्टर, '*daaktar*' (Nepali)

'cycle' = साइकल, '*saaikal*' (Nepali)

However due to literal issues as well as cultural acceptance issues many lexicographers do not favour transliteration. In such case they may coin new multiword expressions like:

'binary' = द्वि_आधारी_अंक '*dwi_aadhaari_ank*' or

'binder' = बाँध्नका_लागि_प्रयोग_गरिने_वस्तु '*baandhnaka_laagi_prayog_garine_wastu*'

as well as new words like:

'pilot' = विमानचालक, '*vimaanchaalak*',

'cycle'= द्विचक्रयान '*dwichakrayaan*',

'rickshaw' = त्रिचक्रयान '*trichakrayaan*'

as a solution to this case.

But in many cases such new words or multiword expressions may not be user friendly and thus will be limited in use in literary works and official works only. Common people more often will use the transliterated forms (in exact or deformed). For example a rickshaw puller or a passenger will prefer to use 'rickshaw' instead of त्रिचक्रयान.

Multiword expressions may get very long. For example: {मठरी, [*maTharee*, an eatable made of wheat flour prepared during the Hindu festival of छट पूजा '*chhaT poojaa*']} is a culture-specific sense in Hindi with no counterpart in Nepali. Transliteration will give मठरी as it is in Nepali. However a possible multiword expression in Nepali can be, 'गहुँको_पीठोले_बनाएको_खाने_कुरा', '*gehũko_piThole_banaaeko_khaane_kuraa*' but it will be quite long, still lacking information about the festival name.

Coining of new word in the target language is also a challenge because this needs synchronization among lexicographers.

In such a case, if the sense will never be required in Nepali context then the creation of the corresponding synset in Nepali may be avoided.

Case c: This is the reverse of Case b. In this case after the manual creation of synset and relation establishment in NWN if we wish to introduce the sense in HWN then we will have to face the same challenges as mentioned for Case b above. Further deciding upon a sense as Nepali-specific needs synchronization among lexicographers of Nepali which is again a challenge.

There are other linguistic challenges also that are faced in the development process like:

1) The attenuated gender system in Nepali raises a peculiar problem. The same word have to be used to mean both genders of non-human animates, like, 'peacock' and 'peahen' both are referred to as 'मुजुर', '*mujur*' in Nepali. Sometimes the difference is established using a prefix 'पोथी', '*pothi*' forming a multiword expression in Nepali for the non-human female animates, like 'पोथी_बाघ', '*pothi_baagh*' for 'tigress. Though for some commonly referred non-human female animates like 'cow' etc. feminine terms like गाइ, '*gaai*' exist, but in general it is not so.

2) When a lexicographer conceptualizes the synset creation process at word level then it may so happen that the lexicographer may find a word in Hindi that has more number of senses in Nepali than in Hindi. In that case a Nepali lexicographer may get tempted to add the additional senses also in the NWN. Then for those uncommon Nepali-specific senses the lexicographer will have to face the problems for Case c.

3) Nepali contains fewer synonyms in comparison to the Hindi synsets for most of the senses.

4   The WordNet creation tool

The development of the NWN is being carried out presently using the Expansion Approach with the help of a browsable-searchable software tool for the WordNets. The software tool provides an interface for each field (discussed later) of the synsets of the HWN and NWN as shown in figure 1. The tool does not allow the fields of HWN to be modified as it is the pivot. It also contains a link (read-only) to refer the English WordNet for reference. The tool has been developed at CFILT, IIT Bombay (http://www.cfilt.iitb.ac.in/). The front-end of the tool has been implemented in Java. The Java interface is connected at the backend with text files of synsets, called "syns" files, for Hindi and English. As the lexicographer inserts corresponding synsets in Nepali against the Hindi ones (sometimes referring the corresponding English ones) the output syns file for the synsets of Nepali is created. The choice of

textual database is for simplicity and to extend support for multiple platforms without the need of installation of any DBMS server like MySQL etc. by the end user. Each synset entry in a "syns" file has five fields:

ID: The synset identifier.
CAT: The syntactic category of the sense.
CONCEPT: It explains the concept represented by the synset. For example, "यस्तो कुरा वा काम जसले कसैको मान वा प्रतिष्ठा कम गराउँछ" (*yasto kuraa waa kaam jasle kasaiko maan waa pratishTha kam garaaũcha*) explains the concept of insult as some saying or deed which diminishes somebody's reputation.
EXAMPLE: It gives the usage of the words of the synsets in the sentence. In general, the words in a synset are replaceable in the sentence. For example: "हामीले कसैलाई पनि अपमान गर्नुहुँदैन" (*haameele kasailaaee pani apmaan garnuhũdain*) gives the usage for the words in the synset of 'अपमान', '*apmaan*' representing insult as something that should not be done to anybody.
SYNSET-(LANGUAGE): It keeps the set of synonyms for the sense in the LANGUAGE designated. In the output syns file for NWN this field has the name SYNSET-NEPALI.

It is important to mention that the synset identifier 'ID' is the key to connect two WordNets. Also for a given polysemous word, for each of the sense of the word, there will be a separate ID.

The tool depicted in figure 1 has an intuitive interface and contains several features ideal for the expansion approach. Features such as searching a synset by ID or by a word, listing all complete, incomplete or all synsets of target language Nepali and font increasing/decreasing for better readability are also there. Provision is also there for adding extra comment, if necessary, with a synset in NWN.

The NWN presently has around 3000 synsets consisting of nouns, verbs, adjectives and adverbs. Since it is currently under development so at different stages in its development phase different numbers of new synsets will get introduced in NWN as such the "syns" text file format for synsets' storage and exchange seems quite ideal. A text file is easy to exchange for purposes like verification and rectification, however such plain text file are always vulnerable to error due to mishandling. Once all the synsets of Hindi gets linked with their Nepali counterparts the set of all error-free 'syns' files can then be coalesced together and stored in a DBMS server like MySQL with proper security implementations for online browsing (Online Hindi WordNet, 2009) or in DBMS like JavaDB (JavaDB, 2009) for embedded systems.
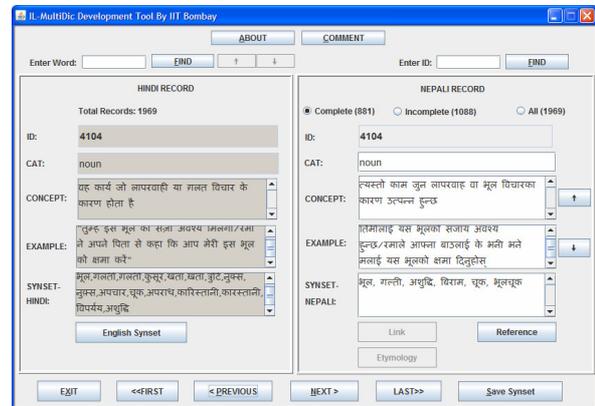

Figure 1: The WordNet creation tool from CFILT, IIT Bombay

5 Conclusion

In this paper we have discussed some characteristic features of Nepali, the expansion approach of Nepali WordNet creation using relation borrowing, linguistic challenges involved, software tool that is in use for the development of Nepali WordNet and the storage structure for WordNet entries.

The expansion approach is very useful considering the time and effort needed in creating WordNets. It avoids duplication of effort. The linguistic challenges discussed in context of Nepali mainly imply the challenge of obtaining a one-to-one correspondence for senses in the Hindi WordNet in to Nepali WordNet.

One of the aims of developing the Nepali WordNet is to overcome the problem of language barrier for the common Nepali speaking peoples who solely use Nepali. When completely implemented, the Nepali WordNet will turn out to be a milestone for the Nepali language. As a future work the authors are interested in implementation of Lesk algorithm and the like for Nepali Word Sense Disambiguation.

Acknowledgements

References

Colin P. Masica. 1991. *The Indo-Aryan Languages*. Cambridge University Press, Cambridge, UK. http://books.google.com/books?id=J3RSHWePhXwC&pg=PA221

Debasri Chakrabarti, Dipak Kumar Narayan, Prabhakar Pandey, and Pushpak Bhattacharyya. 2002. Experiences in building the Indo WordNet - A WordNet for Hindi. 1st Global Wordnet Conference (GWC 02), Mysore, India, January, 2002.

Fellbaum, C. (ed.) 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1993. Five Papers on WordNet. MIT press. http://www.mit.edu/~6.863/spring2009/readings/5papers.pdf

George A. Miller. 1995. English WordNet - A Lexical Database for English. *Communications of the association for Computing Machinery*, 38(11):39-41.

George Cardona and Dhanesh Jain (eds.) 2003. *The Indo-Aryan languages*, volume 2. Routledge Language Family Series, London and New York. http://books.google.com/books?id=jPR2OlbTbdkC&pg=PA554

Hindi WordNet Documentation. 2009. http://www.cfilt.iitb.ac.in/wordnet/webhwn/other/hwn_docs_2.doc.

JavaDB. 2009. http://developers.sun.com/javadb

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

Nepali language. 2009. http://en.wikipedia.org/wiki/Nepali_language.

Online Hindi WordNet. 2009. http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php

Piek Vossen. 2002. EuroWordNet General Document. EuroWordNet Project LE2-4003 & LE4-8328 report, University of Amsterdam.

Zoltán Alexin, János Csirik, András Kocsor, and Márton Miháltz. 2006. Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction. 3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea, January, 2006.