Expansion of the First Hindi-Nepali Word-Net based Bi-Lingual Dictionary and the advancement of the Human-Machine Interface

Shantanu Kar Alok Chakrabarty

Shaansoft Technologies, Silchar Area, India

ABSTRACT

Natural Language Processing is introducing a new era in the field of Computer Science and Machine translation. Human- Machine interaction is to play a very important role in the coming centuries as the dependency of human over the machine is increasing variably. Word-Net was first introduced by Miller and Fellbaum in 1985. WordNet is a Lexical database for the Human Languages. It groups the Human Language into sets of synonyms called "synsets" which provides short, general definitions, and records the various semantic relations between these synonym sets. Word-Net based Bi-Lingual Dictionary (Started in January, 2009) is a part of the Nepali WordNet Project under Indo-WordNet Project. It is an approach towards the construction and application of Multilingual Indo-WordNet, which is a pioneering effort on Nepali to Hindi and vice versa translation. The making of the Hindi-Nepali Word-Net was a challenge to us as the machine has to translate the sense of each word to produce the expected result. One of the major reasons behind this is the non-availability of rich lexical re-sources in Hindi and Nepali. The Hindi-Nepali WordNet based dictionary is thus an endeavor to prepare a rich lexical resource for the Hindi and Nepali Languages for effective machine translation and to facilitate the development of Information and Communication Technologies in Hindi-Nepali. The endeavor is inspired by the famous English WordNet, Hindi Word-Net and the Nepali Word-Net (Nepali Word-Net is also, created by us as a research and project work at Assam University, Silchar as a part of the Indo-Word-Net Project). In this paper we have discussed about the expansion approach of first Hindi-Nepali WordNet based Bi-Lingual Dictionary, the linguistic challenges involved, Word-Net creation tool interface and the synsets' storage structure.

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." Language translation has always been of huge importance for human communication especially when two human beings are of two different mother tongues. In a WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms or synsets, each of which is 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 Hindi-Nepali WordNet based dictionary is an attempt to prepare such a lexical reference system for the languages along the lines of the famous English Word-Net (Fellbaum, 1998; Miller, 1995), Hindi WordNet (Chakrabarti, 2002) and the Nepali Word-Net (we, 2009) so that it can be used as a tool for enhancing the performance of Machine Translation and cross lingual information retrieval systems involving Hindi-Nepali and to facilitate the development of various Information and Communication Technologies.

THE CONTENTS FOR THE REST OF THE PAPER IS AS FOLLOWS:

Section 2, Features of Nepali Translation . Section 3, is on the Construction & Architecture of the advanced WordNet based bilingual Dictionary and the relation borrowing concept. Section 4, presents a discussion on the Overview of the Human-Machine Interface of the WordNet based bilingual Dictionary, Interface and the DFD', and finally Section 5 concludes the paper.

2. FEATURES OF NEPALI TRANSLATION

Nepali (नपाली) is written left to right in the Devanagari script. Nepali goes by various names. It was also called **Gorkhali** or **Gurkhali** (i.e., the language from Gorkha which later gave its name to the famous Gurkhas). Nepali (नपाली) 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: उसल मरो करा खायो। Transliteration: usle mero keraa khaayo. Gloss: he my banana ate. Parts: Subject Object Verb Translation: He ate my banana.

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: कटो आयो, कटी 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. Gen-der in Nepali is a syntactic property. For both the genders a common pluralizing suffix 'हरु', 'haru' can be used for nouns in Nepali, like कटाहरु, '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. CONSTRUCTION & ARCHITECTURE OF THE ADVANCED WORDNET BASED BILINGUAL DICTIONARY

Synset are the basic building blocks of WordNet. In the WordNet the words are grouped together according to their similarity of meanings. Two words that can be interchanged in a context are synonymous in that context. For each word there is a synonym set or Synset, in the WordNet, representing one lexical concept. This is done to remove ambiguity in cases where a single word has multiple meanings. The WordNet deals with the content words, or open class category of words.

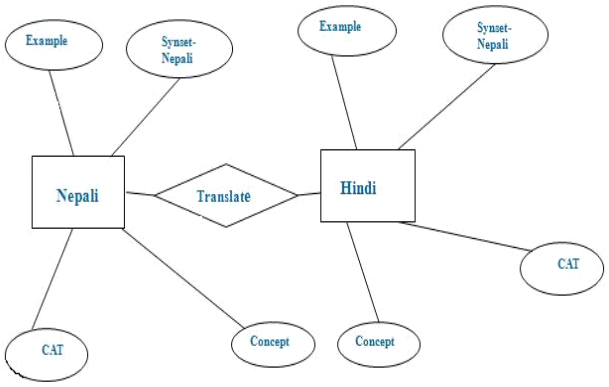
Each entry in the WordNet consists of following format:

CAT: It denotes the category i.e. parts of speech.

CONCEPT: It explains the concept denoted by the Synset.

EXAMPLE: It gives the usage of the words in the sentence.

SYNSET-HINDI/NEPALI: It is a set of synonymous words.



In this project we modified the previous version of the Word-Net based dictionary keeping in mind the popular words which are mostly used in both Nepali and Hindi and display them in such a manner so that user can easily understand it. For the user to recognize a word easily we included the concept of the word, a sentence to describe the use of the word and its synonyms. Later we recognize that all users may not know how to type Nepali or Hindi alphabets and to overcome this difficulty we developed a "Devanagari Keyboard".

The challenges we faced in the construction of the first WordNet based Bilingual Dictionary has been solved. But the new challenges arise. For example: The word "आम, **Amm**" may refer to "**Amm Admi** (Common Man)" or the Fruit "आम (Mango)". The challenge is to pick the correct sense of the word, if we are to translate the word.

NEPALI-HINDI TRANSLATION DICTIONARY:

In an actual Machine Translation situation, for every word in the source language a single word or phrase in the target language will have to be produced.We have collected the Nepali words and then we put them in the following Synset form:

CAT :: Noun

CONCEPT :: मित्रावरूणले उर्वशीलाई देरख्दा पतन भएको वीर्य कुम्भमा पर्दा जन्मेको भनिने एक ऋषि,गोत्रप्रवर्तक अनेक ऋषिमध्ये एक ऋषि,पुराणमा वर्णित कथानुसार समुद्रको सम्पूर्ण जल पान गर्ने ऋषि। धिचुवा ऋषि, खन्चुवा ऋषि।

EXAMPLE:: कस्तो अगस्ति रहेछ सम्पूर्ण जहानलाई पकाएको खाना एक्लैले घिच्यो।

अगस्ति विन्ध्याचल पर्वतका गुरु थिए।

SYNSET-NEPALI :: अगस्ति, अगस्त्य, कुम्भयोनि, एकतारा, अगस्तितारो

We also have collected the Hindi words and then we put them in the following Synset form:

CAT :: Noun

CONCEPT :: वह जो धनुष धारण करता हो **EXAMPLE ::** ऋषि ने देखा कि साधु वेश में दो धनुषधारी वन में विचरण कर रहे हैं

SYNSET-HINDI:: धनुषधार,धनुर्धार, धनुर्धर,धनुर्द्धर,धनुर्द्धर,तीरंदाज,तीरंदाज, बानैत,शारंगधार,धनुर्भृत,धनुर्ग्रह |

Then we selected the common Nepali and Hindi words from the Nepali Synset file and the Hindi Synset file and then we linked them using Java code and inserting them in the MySQL database. Finally we linked those data to our GUI interface using Java Database Connectivity (JDBC).

4. OVERVIEW OF THE HUMAN-MACHINE INTERFACE OF THE WORDNET BASED BILINGUAL DICTIONARY

WordNet based Bilingual Dictionary is divided into two modules design, simplicity and flexibility. The modules are :

I. HINDI WORD INFORMATION:

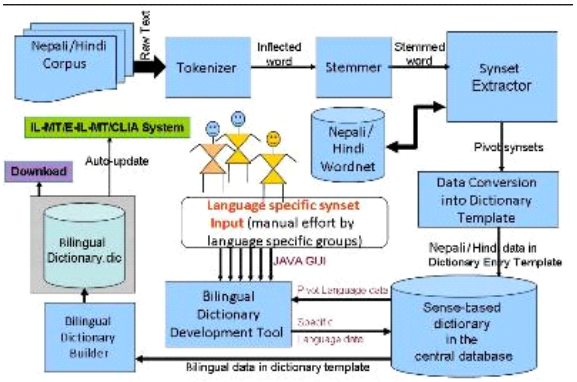
It contains a table containing Hindi words, meaning of the words and examples to describe the use of those words.

II. NEPALI WORD INFORMATION:

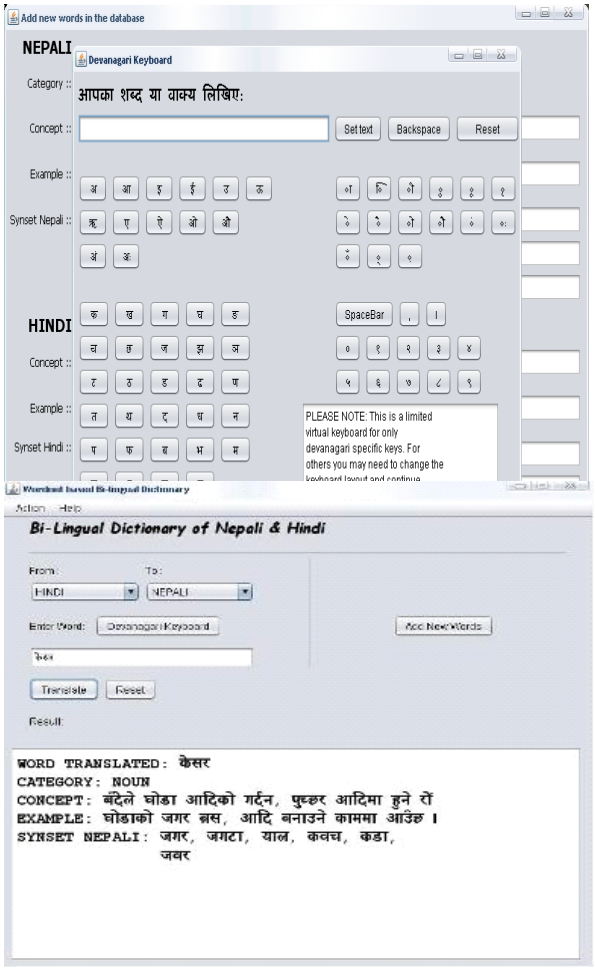
It contains a table containing Nepali words, meaning of the words and examples to describe the use of those words.

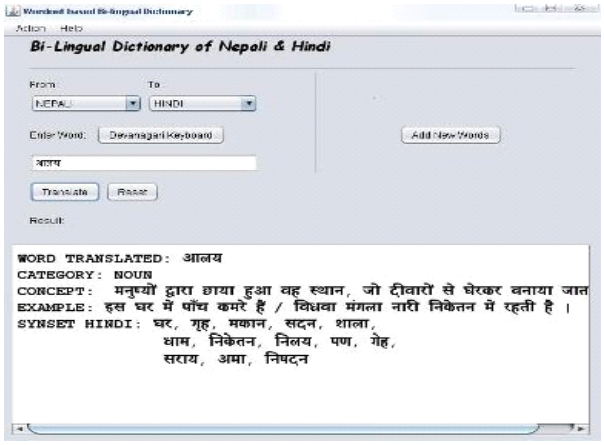
This two different databases of Hindi and Nepali languages are interlinked by the language translation algorithm corresponding to the ID specified to each of the distinct words in Hindi and Nepali.

THE DFD IS AS FOLLOWS:



INTERFACE OVERVIEW:





5. CONCLUTION

More than 50% of the Indian population speaks Hindi. This Word-Net based dictionary will help people knowing Hindi to

learn Nepali also the people knowing Nepali to learn Hindi. Moreover it is the first approach to interlink the Indian languages, which will be used for the language translation purpose in future.

In this paper we have discussed some characteristic features of Nepali and Hindi, the expansion approach of Hindi-Nepali WordNet based bilingual dictionary using relation borrowing, linguistic challenges involved, software tool that is in use for the development of Hindi and Nepali WordNet and the storage structure for WordNet entries. The expansion approach is very useful considering the time and effort needed in creating the WordNet. 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 and vice versa.

6. ACKNOWLEDGEMENTS

This research and development was supported by a grant from the Department of IT, Ministry of Communication and Information Technology, Govt. of India. We acknowledge the WordNet Group at IIT Bombay for their support and specially Prof. Pushpak Bhattacharyya, Consortium Leader, NE WordNet project and Prof. Bipul Shyam Purkayastha,Chief Nepali WordNet project, Assam University and for his constant encouragement and support. Further we acknowledge the efforts of Dr.Khagen Sharma, Linguist, Nepali WordNet group and Mr.T.N.Upadhaya, Linguist & Lexicographer, Nepali WordNet group, in the preparation of this paper. We also acknowledge Mr. Debasish Roy, Ms. Baisakhi Dey, Ms. Mallika Gogoi for the tremendous support in the construction of the World's First Hindi-Nepali WordNet based Bi-lingual Dictionary.Finally we convey our thanks to the Co-workers at Shaansoft Technologies for the help and support offered.

7. REFERENCES

[1] Shantanu Kar and Alok Chakrabarty, "An Approach Towards The Construction Of The First Hindi-Nepali Word-Net Based Bi-Lingual Dictionary And The Challenges Handled," in proceedings of the ICACTEA'11: 1st IEEE International Conference On Adaptive Computer Technologies in various Engineering Applications, 2011.

[2] Alok Chakrabarty, Bipul Syam Purkayastha "Experiences in building the Nepali WordNet - insights and challenges, 5th Global Wordnet Conference (GWC 10), IIT Bombay-2010"

[3] 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.

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

[5] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller.1993. Five Papers on WordNet. MIT press.

[6] George Cardona and Dhanesh Jain (eds.) 2003. The Indo-Aryan languages, volume 2. Routledge Language Family Series, London and New York.

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

[8] Manish Sinha, Mahesh Reddy, Pushpak Bhattacharyya. 2006. An Approach towards Construc-tion and Application of Multilingual IndoWordNet. 3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea, January, 2006.

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

[10] Nepali language. 2009. http://en.wikipedia.org/wiki/Nepali_language. Online Hindi WordNet. 2009.