Towards Building Advanced Natural Language Applications - An Overview of the Existing Primary Resources and Applications in Nepali

Bal Krishna Bal
Madan Puraskar Pustakalaya
Lalitpur,Patan Dhoka,
Nepal
bal@mpp.org.np

Abstract

The paper gives an overview of some of the major primary resources and applications developed in the field of Natural Language Processing(NLP) for the Nepali language and their prospective for building advanced NLP applications. The paper also sheds light on the approaches followed by the current applications and their coverage as well as limitations.

1 Introduction

NLP is a relatively new area of involvement in the context of Nepal.The first ever NLP works in Nepal include the Nepali Spell Checker and Thesaurus that got released in the year 2005. The years after that saw an increasing amount of Research and Development of NLP resources and applications under different programs.This included **Dobhase**[1], an English to Nepali Machine Translation System, Stemmer and Morphological Analyzer, Parts-of-Speech(POS) Tagger, Chunker, Parser, a corpus-based on-line Nepali monolingual dictionary, Text-To-Speech etc. On the resources front, by 2008, we have had developed a Lexicon, Nepali Written Corpus,Parallel Corpus, POS Tagset,Speech Recordings etc. In the sections that follow, we will be discussing over the current achievements and the possible advanced applications that can be developed on the basis of the existing resources and applications.

2 Resources

2.1 Nepali Lexicon

The process of the development of the Nepali Lexicon(Bista *et al.*,2004-2007;2004-2007a) underwent several changes as the purpose of the lexicon was not very clear in the beginning.No doubt, we were aware that the usage of a lexicon would be basically a multi-purpose one, fitting to one or more NLP applications but we were a bit unsure about the actual format of the lexicon.As a result, the entries of the lexicon were maintained in different file formats ranging from plain spreadsheets to XML formatted data files. In the beginning, the attributes of the lexicon were decided to be as:

- Rootword
- Headword
- Pronunciation
- Syllablebreak
- Meaning
- PartsofSpeech
- Synonym
- Idiom

But later on, owing to time constraints and also taking into consideration the applicability of the lexicon to some of the immediate NLP applications being developed like the Spell Checker and the Stemmer/Morphological Analyzer, the entries were just made for the attributes - Rootword and PartsofSpeech. The file format was also fixed for the plain spreadsheet one, keeping into consideration the discomfort faced by the linguists and data entry persons with the XML data format.The latest size of the lexicon is 37,000 root words with their parts of speech category specified. Wherever more than one category is possible, multiple categories have been entered with the comma as the separator.

2.2 Nepali POS Tagset

The Nepali POS Tagset designed in the beginning consisted of 112 tags[2]. These tags were

[1] http://nlp.ku.edu.np/cgi-bin/dobhase

[2] http://www.bhashasanchar.org/pdfs/nelralec-wp-tagset.pdf

Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 165–170,
Suntec, Singapore, 6-7 August 2009.

used to manually and semi-automatically annotate the written corpus as well. Experiences, however, showed that error rates of annotation could be much higher when the size of the tagset was a big one, the reason primarily being the chances of assigning incorrect tags to the words out of confusion while manually annotating the the training data itself. It was with such motivations that a smaller sized POS Tagset[3] was later on developed that consists of just 43 tags. While developing the tagset, maximum care has been taken to ensure that this minimalist approach does not unnecessarily eliminate the unavoidable lexical categories of the language. The design of this Nepali POS Tagset was inspired by the PENN Treebank POS Tagset.Hence, whenever possible, the same naming convention has been used as in the case of the PENN Treebank Tagset. In Table 1, we provide the summary of the small sized POS Tagset. Owing to space constraints, we have just provided limited examples for each of the POS category.

2.3 Nepali Written Corpus

Different efforts have been put into developing the Nepali Written Corpus. We try to provide a brief overview on each of them below.

With an aim to facilitate linguistic and computational/corpus linguistic researches of the Nepali language,the compilation of different text corpora got initiated under the activity, **Nepali National Corpus**[4]. The Nepali National Corpus basically consists of three different types of corpora, namely, the Written Corpus(monolingual and parallel), the Spoken Corpus and the Speech Corpus. The monolingual Nepali Written Corpus was further sub-divided into two types - the Core Corpus and the General Corpus. The Core Corpus consists of 398 texts from about 15 different genres and amounting to 1 million words, has been collected from different books,journals,magazines and newspapers. The texts belong to the time period between 1990 and 1992. The Core Corpus has been converted into XML file format. In addition, the text has been annotated using the 112 Tagset. The General Corpus,on the other hand, is a collection of the written texts basically from the

Table 1: Summary of the Nepali POS Tagset

Tag	Description	Example
NN	Common Noun	Ghar
NNP	Proper Noun	Ram
PP	Personal Pronoun	Ma
PP$	Possessive Pronoun	Mero
PPR	Reflexive Pronoun	Afu
DM	Marked Demonstrative	Arko
DUM	Unmarked Demonstrative	Tyo
VBF	Finite Verb	Khayo
VBX	Auxiliary Verb	Thiyo
VBI	Verb Infinitive	Khana
VBNE	Prospective Participle	hidne manchhe
VBKO	Aspectual Participle	Thiyo
VBO	Other Participle Verb	Diyeko
JJ	Normal Unmarked Adjective	Asal
JJM	Marked Adjective	Ramro
JJD	Degree Adjective	Adhiktar
RBM	Manner Adverb	dhilo hidchha
RBO	Other Adverb	yaha basa
INTF	Intensifier	dherai chalaakh
PLE	Le-Postposition	Harile
PLAI	Lai-Postposition	Bhailai
KO	KO-Postposition	Ramko
POP	Other PostPositions	tabulmathi
CC	Co-ordinating Conjunction	ra
CS	Subordinating Conjunction	Kinabhane
UH	Interjection	Oho
CD	Cardinal Number	Ek
OD	Ordinal Number	Pahilo
HRU	Plural Marker	Haru
QW	Question Word	Ko
CL	Classifier	Dasjana
RP	Particle	Khai
DT	Determiner	Tyo keto
UNW	Unknown Word	Nekomprenas
FW	Foreign Word	good
YF	Sentence Final	? ! etc.
YM	Sentence Medieval	, ; : etc.
YQ	Quotation	'' ""
YB	Brackets	() {} []
FB	Abbreviation	Ma.Pu.Pu
ALPH	Header List	Ka.
SYM	Symbol	%
NULL	<NULL>	

[3] http://nepalinux.org/downloads/nlp/nepali_pos_tagset.pdf

[4] The Nepali National Corpus was developed under the Nepali Language Resources and Localization for Education and Communication(NeLRaLEC)Project.For details, please visit http://bhashasanchar.org

internet. The collected texts amount to a size of 14 million words. Both the Core Corpus and the General Corpus above have been developed following the internationally accepted FLOB and FROWN framework for collecting text corpus.

Another set of collection under the Written Corpus is the Parallel Corpus. The Parallel Corpus consists of collections from two genres - computing and national development. The one on computing sizes to be 3 million words of English-Nepali parallel texts whereas the other one on national development amounts to about 966,203 words.

In another bid, a Nepali Corpora parallel to 100,000 words of common English source from PENN Treebank corpus, available through the Linguistic Data Consortium(LDC)has been developed[5].This Parallel Corpus has been also POS Tagged with the 43 POS Tagset as presented in Table 1.

2.4 Nepali Spoken Corpus

The Spoken Corpora has been designed on the basis of the Goteborg Spoken Language Corpus(GSLC). The Corpora have been collected from 17 social activities and contain about 2,60,000 words. These texts are audio-video recordings of the activities with their corresponding transcriptions and annotations about the participant's information. Each activity is stored in three separate files (.mpeg, .txt and .doc) respectively for recording,transcription and recording information.

2.5 Nepali Speech Corpus

The Speech Corpus is a specialized recordings of speech developed for the Nepali Text-To-Speech(TTS) application for enabling the software to speak Nepali from written texts.It consists of 1,880 sentences and 6053 words, extracted from the Core Corpus and later recorded in male and female voices. The recordings are approximately of 3-4 hours.

3 Applications

3.1 Nepali Thesaurus

For developing the Nepali Thesaurus, we have used the MyThes framework[6] developed by Kevin Hendricks. MyThes is incorporated with the OpenOffice.org suite.Originally, it did not support UTF-8 encoding but the support has been enabled after OpenOffice.org 2.0.2 onwards.The Nepali Thesaurus currently contains 5,500 entries with the attributes - POS Tag, Meaning and Synonym. This application has been released as an inbuilt package with OpenOffice.org Writer localized into Nepali for public usage since 2005.

3.2 Nepali Spell Checker

The Nepali Spell Checker follows the Hunspell framework[7]. In essence, Hunspell is a spell checker and morphological library.It is included in OpenOffice.org suite 2.0 onwards by default. Recently,it has also been adopted by the Google Search Engine as it's default Spell Checker. Depending upon the language specific terriotory, HunSpell may be customized by using the concerned locale file. HunSpell requires two files[8], respectively the dictionary file that contains the words for the language and the affix file that has rules associated to the words in the dictionary by using flags serving as pointers. The two files should be located in the folder openofficefolder/share/dic/ooo/. Spell checking is done using the affix file, locale and the dictionary file.While the affix file consists of affix rules, the dictionary file consists of root words. At the moment,the size of the dictionary file is about 37,000 entries whereas we have about 1,800 affix rules in the affix file. The word coverage in terms of spell checking is 6.2 million Nepali words. Random tests of the spell checker yielded accuracies of 90%(43 words unhandled out of 450 words),94%(25 words unhandled out of 400 words),89% accuracy(100 words unhandled out of 923 words) etc. By saying unhandled, we refer to the situation whereby the incorrect words are not provided appropriate suggestions.

3.3 Dobhase - English to Nepali Machine Translation System

With a view to aid to the majority of English unproficient Nepalis to some extent, this application was developed under a joint collaboration between Kathmandu University and Madan Puraskar Pustakalaya.The software currently is able to provide gist translations to simple declarative sentences.It

[5]This work has been developed with the support from the Language Resource Association (GSK) of Japan and International Development Research Center (IDRC) of Canada, through the PAN Localization Project(www.PANL10n.net)

[6]http://lingucomponent.openoffice.org/MyThes-1.zip

[7]http://hunspell.sourceforge.net

[8]The two files for Nepali is available at http://nepalinux.org/downloads/ne_NP_dict.zip

is a rule and transfer-based Machine Translation System. More information on the software is available at http://nlp.ku.edu.np/

3.4 The Online Nepali Dictionary

A Corpus based Online Nepali Dictionary has been developed for Nepali and is available in the following link http://nepalisabdakos.com
This dictionary differs from the existing ones(both hard and soft copy versions) in that this dictionary contains examples and meanings from the corpus itself. The XIARA software has been used to look for wordlists and concordances in due course of compiling the dictionary. The dictionary currently contains about 8,000 entries .

3.5 The Nepali Text-To-Speech

The Nepali Text-To-Speech Application has been developed following the Festival Speech Synthesis System.Currently, the Nepali Text-To-Speech works just in the Linux environment and has the basic capabilities of reading text from files.One may opt to hear the texts either from a male or a female voice.The application has been deemed useful not only to visually impaired but also to illiterates.Lately, there has been a growing demand of the application for extending it to a screen reader and making it work in cross-platforms.For more information, please visit http://bhashasanchar.org/textspeech_intro.php

3.6 Conversion Tools

Keeping into consideration that a lot of texts both in the government and the general public are still encoded in ASCII-based Nepali non-unicode fonts,we have developed the **Conversion Tools** both for converting non-unicode texts to unicode and vice versa.For details,please visit http://madanpuraskar.org/ Our efforts in the development of the tool have been supplemented by the Open Source Community as well.The latest information on the extended work is available at http://code.google.com/p/nepaliconverter/

3.7 The Nepali Stemmer and the Morphological Analyzer

The Nepali Stemmer and the Morphological Analyzer combines the results of the Stemmer and the Morphological Analyzer in the sense that besides producing the stem or root of any word, the associated bound morphemes and their grammatical category are also kept track of. The Nepali Stemmer and Morphological Analyzer is a rule-based one and makes use of the following resources:

- Free morpheme based lexicon
- Affix file or bound morpheme list
- Database of word breaking rules

The free morpheme based lexicon consists of free or unbound morphemes of the Nepali language together with their respective parts-of-speech information. Similarly, the affix file or bound morpheme list contains the prefix and the suffixes in Nepali.These affixes are further associated with numbers which point to the corresponding word breaking rules.Finally, the word breaking rules database basically represent the insertion and deletion rules applicable once a word breaks down into the root and the respective affixes. The application, which is still at a prototypical stage, is available at http://nepalinux.org/downloads/nlp/stemmer_ma.zip

3.8 The Nepali Computational Grammar Analyzer

The Nepali Computational Grammar Analyzer is an attempt to develop a basic computational framework for analyzing the correctness of a given input sentence in the Nepali language. While the primary objective remains in building such a framework, the secondary objective lies in developing intermediate standalone NLP modules like the POS Tagger,chunker and the parser.In Figure 1, we present the system architecture of the Nepali Computational Grammar Analyzer. Talking about the individual modules,for the POS Tagger,we have used TNT[9],a very efficient and state-of-the-art statistical POS tagger and trained it with around 82000 Nepali words.Currently the accuracy of the trained TnT POS Tagger for Nepali is 56% for unknown words and 97% for known words.
Similarly,for the chunker module, we have developed a hand-crafted linguistic chunk rules and a simple algorithm to process these rules.Currently, we have around 30 chunk rules, which have to be further optimized for better coverage and output.The chunkset consists of 11 chunk tags at the moment.
For the parser module, we have implemented

[9] http://coli.uni-saarland.de/ thorsten/tnt/

a constraint-based parser following the dependency grammar formalism[10] and in particular the Paninian Grammar framework(Bharati *et al.*,1993,1995;Pederson *et al.*,2004).A dependency parser gives us a framework to identify the relations between the verb(s) and the other constituents in a sentence. Such relations, which basically occur between verb(s) and nominal constituents are called **Karaka relations**.For Nepali, we have identified altogether six different Karaka relations, namely,Karta - K1, Karma - K2, Karan-K3, Sampradan - K4, Apadaan-K5, Others-KX.

The assumption is that if we can establish valid Karaka relations between the chunks of the sentence and the verb, then the given input sentence is valid.For example, in the Nepali sentence *Ram le bhaat khaayo (meaning Ram ate rice)*, there is a K1 relation between the verb *khaayo* and the noun chunk *Ram le*, and similarly K2 relation between the verb *khaayo* and the noun chunk *bhaat*. Next, the Karaka frame specifies what karakas are mandatory or optional for the verb and what vibhaktis(postpositions) they take respectively. Each verb belongs to a specific verb class (in our case, whether the verb is transitive or intransitive)and each class has a basic karaka frame. Each Tense,Aspect and Modality - TAM of a verb specifies a transformation rule. Based on the TAM of a verb, a transformation is made on the verb frame taking reference of the TAM frame. A detailed description of the Nepali Computation Grammar Analyzer is available at http://nepalinux.org/downloads/nlp/report_on_nepali_computational_grammar.pdf

The Grammar Analyzer for Nepali currently parses and analyzes simple declarative sentences with just one verb candidate.We have developed the karaka frame for around 700 Nepali verbs.An agreement module if added to the analyzer could further filter the parses returned by the parser module, this time taking the feature agreements like gender, number, person, tense, aspect and modality. Hence,with the possible addition of the agreement module, the robustness of the Grammar Analyzer is believed to significantly increase.

[10]http://w3.msi.vxu.se/nivre/papers/05133.pdf

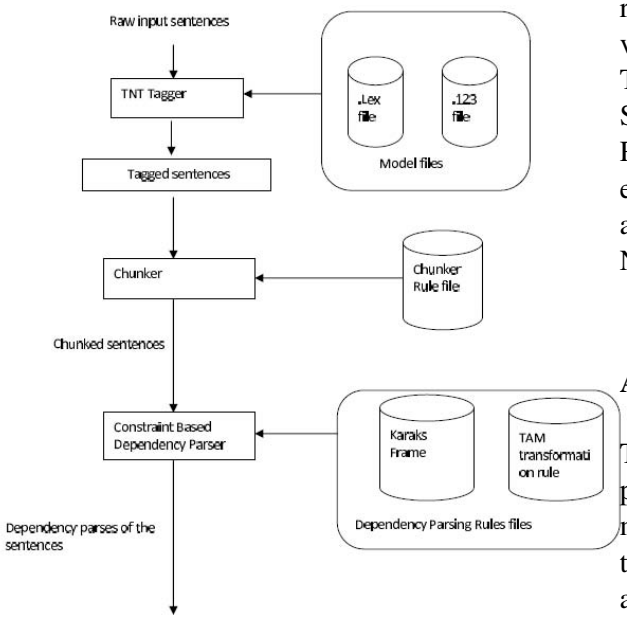

Figure 1: System Architecture of the Nepali Computational Grammar Analyzer

4 Conclusion

In this paper, we discussed on the different efforts put towards developing the basic NLP resources and applications.We also talked about the approaches followed by the applications and shed light on their current coverage and limitations. From the discussions above, it is quite clear that much work has been done in developing a basic NLP foundation for Nepali both from resource buiding and applications development perspectives. The way ahead is undoubtedly in refining the current achievements and building advanced NLP applications like Statistical Machine Translation System, Name Entity Recognition System, Question Answering System,Information Retrieval System, Information Extraction System etc. Another possibility is applying the expertise and experiences gathered while working for Nepali to other non-Nepali languages.

Acknowledgements

The works described in the paper have been partly supported by the International Development Research Center(IDRC), Canada through the PAN Localization Project(http://panl10n.net) and Asia IT & C Programme of the European Commission under the Bhasha Sanchar Project(http://bhashasanchar.org).

References

A.Bharati, V. Chaitanya, and R. Sangal, Natural Language Processing - A Paninian Perspective. New Delhi: Easter Economy Edition ed.Kantipur:Prentice Hall, 1995.

A.Bharati and R. Sangal, "Parsing free word order languages in the Paninian framework.," in Proceedings of the 31'st Annual Meeting on Association For Computational Linguistics (Columbus, Ohio, June 22, 1993).Annual Meeting of the ACL., Morristown, NJ, 1993, pp. 105-111.

A.Bharati, R. Sangal, and T. Reddy, "A Constraint Based Parser Using Integer Programming," in Proceedings of the ICON-2002, Mumbai, 2002, pp. 121-127.

B. K. Bal, B. Karki, and L. Khatiwada, "Nepali Spellchecker 1.1 and the Thesaurus, Research and Development," PAN Localization Working Papers 2004-2007.

B. K. Bal and P. Shrestha, "Nepali Spellchecker," PAN Localization Working Papers 2004-2007.

S. Bista, L. Kathiwada, and B. Keshari, "Nepali Lexicon," PAN Localization, Working Papers 2004-2007, pp.307-10.

S. Bista, L. Khatiwada, and B. Keshari, "Nepali Lexicon Development.," PAN Localization, Working Papers 2004-2007, pp.311-15.

M.Pederson, D. Eades, S. Amin, and L. Prakash, "Relative clauses in Hindi and Arabic:A paninian dependency grammar analysis. ," in Proceedings of the Twentiet International Conference on Computational Linguistics., Geneva, 2004, pp. 17-24.