Nepali Spellchecker

Bal Krishna Bal, Prajol Shrestha
Madan Puraskar Pustakalaya
bal@mpp.org.np,prajolshrestha@gmail.com

Abstract

The Nepali Language is in itself a complex language owing to a lot of inflections and derivations in word formation. Writing in Nepali correctly in terms of accepted spelling and grammar norms is a problem sometimes even to experts. In this context, making the machine check spelling errors provided some limited grammatical knowledge is still more challenging.

With an aim to integrate Spell-Checker for Nepali in OpenOffice.Org, the Development and Linguist team at Madan Puraskar Pustakalaya, Nepal conducted a research work and on the basis of that implemented a basic Nepali Spell-Checker. Currently, the word coverage of the Spell-Checker is approximately 300,000 words. This document is a collection of experiences gained by the Madan Puraskar Pustakalaya Team in due course of the implementation of the Spell-Checker for Nepali in OpenOffice.org.

1. Introduction

A Spell checker, in general, is an application that aids end-users in correctly writing texts in text editors. The general mechanism devised in Spell checking applications comprises of the following:

- Certain words are written by the end user in some text editors.
- The Spell Checker, incorporated in the text editor would then check the correctness of the word(s) being written thus highlighting the misspelled words if any. In addition, the application also facilitates possible suggestions for the incorrect words.

Integrating the Spell Checker Application for Nepali in OpenOffice.org as for any other language has the following requirements:

- The Application has to be written as according to the OpenOffice.org framework. This involves the usage of the format and programming structure as provided by the default Spell Checker in OpenOffice.org.
- Alternatively, a standalone application could be developed, which could be later incorporated into OpenOffice.org using the Unified Network Component (UNO) which does the work of a supporting component for making possible communication between the spell checker application and OpenOffice.org.

The available spell checkers in OpenOffice.org are respectively, myspell,Ispell ,which have no unicode support and aspell, hunspell,which have unicode support. All four of these applications share almost the same file formats with the only difference that only hunspell has a better support for Unicode, which is a must for languages like Nepali which follows complex script (Devanagri) and hence requiring multi-byte encoding for a single character.

2. Methods

Nepali language being complex in it's form, has almost 75% of its general vocabulary comprising of derived and inflectional words rather than head words. This has certainly need to be taken into consideration while developing the Spell Checker given the fact that it is almost impossible to include all the derived and inflectional words together with the head words and on the other side without their inclusion, the Spell Checker would be incomplete. This problem can be effectively addressed by creating affix rules that is rules for prefix, suffix and infixes for the head words. Some of the examples of prefixes and suffixes being used in Nepali words are given below:

(अ)सहमति -अ prefix सहमति head word

(अ)समान -अ prefix समान head word

खा(नु) -नु suffix खा head word

गर्(नु) -नु suffix गर् head word

पढ्(नु) -नु suffix पढ् head word

रू(नु) -नु suffix रू head word

धु(नु) -नु suffix धु head word

While evident from the above examples to some extent, the regular pattern of word formation suggest that generally the nouns take the prefix अ to form compound words and that verbs take the suffix नु thus resulting into new words. Such word formation pattern may be exploited for creating effective affix rules. The creation of affix rules to be later associated with head words substantially frees oneself from exhaustively include all the possible words of the language in the lexicon or the dictionary for spell checking purpose but at the same time has its own dark side. The major weakness of the affix rules approach is the amount of time and correspondingly computational resources (processor speed) that is required for checking whether a certain affix rule applies to a word and then later forming the compound word on the basis of the rules so as to check the correctness of the word written by the end user in the text editor. In a comparatively slow and a computer with low memory, the process might turn up to be rather slack and disgusting. But given the time required for the Spell Checker development and the potentiality of wide word coverage, this option still seems to be appealing.

3. Discussion: Implementation of the Spell Checker in OpenOffice.org

There are two prerequisite files for the spell checker in OpenOffice.org, namely the Head word file consisting of the head words and the affix files containing the affix rules to be associated with the head words. A simple example dealing with compound word formation via the combination of suffixes is shown below:

Head word file:

2

खा/1

धु/1

Affix file:

SET UTF-8
FLAG num
SFX 0 नु .

The head word file consists of the head word in its purest origin with the rule indicator separated by "/".The number '2' in the head word file represents the number of words in the file while the "SET UTF-8" in the affix file indicates the encoding.

Comparatively for English and some other languages, the alphabets of the language are used to index the affix rules, the latter being less in number but in case of Nepali, owing to their large number numbers should be devised for indexing instead. For this purpose the "FLAG num" is used in the affix file. This will then allow the rules of the affix file to be indexed in numbers. The notation SFX stands for suffix and "0" or zero denotes that nothing is to be deleted and नु is what is to be added to the head word and "." or dot is a regular expression indicating any character the word ends with.

Another important aspect of the spell checking application is the suggestions it has to provide for the wrongly spelt words. There are also inbuilt provisions for this in the spell checker applications available under the OpenOffice.org frame work. The illustration below attempts to describe the procedure:

REP 2
REP ि ी
REP ी ि

Here the notation "REP" stands for replace and would facilitate the suggestion list replacing the ि with ी in words and vice versa. The numeral 2 indicates that there are two such replacements. For example, if somebody wrote नेपालि by mistake, नेपाली would automatically come in the suggestion list. This not only saves computational resources like memory and time but also potentially provides suggestions that rules alone would sometimes not cover.

4. Results

The combined effort of the technical and linguist team of Madan Puraskar Pustakalaya, PatanDhoka, Nepal has come up with a basic Spell Checker for Nepali. As mentioned earlier in the abstract, currently the word coverage is just 300,000 words. The spell checker contains 24,000 head words with 150 affix rules. At a time when the resources for Nepali computing are still in the infancy stage, this could serve as a good initiative for further works in this direction.

5. Conclusion

The current Spell Checker for Nepali though lacks completeness is no doubt a very laudable work in the Nepalese language computing. There are plans for refining this Nepali Spell Checker with head words as much as 50,000 and 750 affix rules in the near future. The work could serve as a baseline for other languages following the Devanagri script.

6. References

For preparing this document the following links and documents were referred to:

[1] "Hunspell"
http://www.sourceforge.net/projects/hunspell

[2] S.K. Bista, B. Keshari, L.P, Khatiwada, P. Chitrakar, S. Gurang, Nepali Lexicon Development, *Forthcoming*