

Nepali Spell Checker

Bal Krishna Bal, Bineeta Pandey, Laxmi Khatiwada, Prajwal Rupakheti

Madan Puraskar Pustakalaya, Lalitpur, Nepal

bal@mpp.org.np, bineeta@mpp.org.np, lkhatiwada@mpp.org.np, prajwal@mpp.org.np

Abstract

Ever since the release of the Nepali Spell Checker version 1.0 together with NepaLinux 1.0 in December 2005, there has been continual works in it's enhancements, the attempts being directed towards making it more robust and of industrial strength. This report gives an overview of the the linguistic and technical efforts made to achieve the target. A comparative analysis of the work performance of the current and previous spell checkers is also given.

1. Introduction

The availability of the Spell Checker in the NepaLinux versions has been a major attraction to the general public. Especially, it has been noted to be of marked help to publication houses, writers and journalists. While the availability of this utility has been applauded, the general feedback has been that it had to be enhanced and provided industrial strength in order to make it's applicability full fledged. It is with such motivations that research and development works on the Nepali Spell Checker were continued. Coming up to the current version, i.e., 3.0 from the first version in 2005, the performance measure of the Spell Checker has greatly improved. This is justified by the word coverage of the current

system which is approximately 6.2 million. The very first version had the word coverage of merely 3,00,000 words.

2. Methods

Enhancing the existing Spell Checkers meant analyzing the shortcomings of the system and fixing them. For this purpose, we have basically adopted the rigorous testing and correspondingly making corrections to the system. In order to test the performance of the system, we collected various Nepali texts from a number of sources, primarily the Internet. These texts were copy pasted in the text editor, OpenOffice.org Writer and the Spell Checker was run. Our focus was on analyzing the words underlined in red throughout the text. Our analysis showed that the underlining in red was displayed mainly due to two reasons – one was because we did not have that particular head word or compound in the dictionary file and the other cause was if it was a suffixed or prefixed word, we did not have the appropriate rules for handling them.

Hence, both head words and compound words detected as being not in the dictionary file were added. The other category of errors were found to be existing because of the improper definition of

the rules and their association with the head and compound words. This was also fixed in the latest system.

3. Discussion

Since almost 75% of the words in the Nepali vocabulary are inflected and derived words, we have strived towards capturing these words with the help of affix rules. Research findings showed that there are four basic types of verb patterns in Nepali, which would cover the most inflections and derivations. For example, there are four basic forms of the root verb "पल्ट": 1. "पल्ट" 2. "पल्टि" 3. "पल्टाउ" and 4. "पल्टाइ". Each of the above form of verbs, generally take the same affixes to generate other additional forms of root verb "पल्ट". For example, when the suffix "नु" is added to all four basic forms of "पल्ट" the words "पल्टनु", "पल्टिनु", "पल्टाउनु" and "पल्टाइनु" are generated respectively. Hence to realize the above, the application of rules is done in two levels: Root verb "पल्ट" is listed in dictionary file and affix rule r1 is applied to it as follows:

पल्ट/r1

Rule r1 contains 4 sub-rules that generates four inflected forms of पल्ट.

SFX r1 ० ०/r2X .

SFX r1 ० ि/r2X .

SFX r1 ० ाउ/r2X .

SFX r1 ० ाइ/r2X .

Rule r1 shown above can generate 4 inflected forms of head word "पल्ट" which are "पल्ट",

"पल्टि", "पल्टाउ" and "पल्टाइ". In the second level, rule r2 is applied to each of the inflected forms as shown below:

SFX r2 ० नु .

SFX r2 ० ने .

SFX r2 ० न्छु [उ]

SFX r2 ० ँछु [उ]

SFX r2 ० यौ [इ]

SFX r2 ० ्यौ [उइ]

Each sub-rule of rule r2 is applied to each of the words "पल्ट", "पल्टि", "पल्टाउ" and "पल्टाइ" provided they end with the character as specified by the regular expression defined at the end of each sub-rule. Since there is a "." at the end of first sub-rule of rule r2, this implies it's applicability to any word. In the given case, it is applied to all four words "पल्ट", "पल्टि", "पल्टाउ" and "पल्टाइ" generating respectively "पल्टनु", "पल्टिनु", "पल्टाउनु" and "पल्टाइनु". However the third sub-rule is applicable to only those words which do not end in "उ". Therefore this sub-rule is applied only to "पल्टनु", "पल्टिनु" and "पल्टाइनु" generating "पल्टन्छु", "पल्टिन्छु" and "पल्टाइन्छु". This sub-rule is not applied to "पल्टाउ", thus preventing "पल्टाउन्छु" from being generated. The fourth sub-rule applies to only those words ending in "उ". Hence it is applied to "पल्टाउ" to generate "पल्टाउँछु".

Rule r1 can generate 4 inflected forms of a single root verb of dictionary. From each of these four verbs, approximately 80 more inflected verbs could be generated after application of rule r2. Therefore, a total of 320(approx) inflected verbs

could be generated from a single verb like “पल्ट” of the dictionary file.

There exists an interesting lexical category in Nepali, which is the postposition. Generally, postpositions come after nouns and sometimes with verbs. So that our rules for postpositions did not necessarily generate inapplicable words (not all nouns and verbs take postpositions), we similarly categorized nouns into five categories. These categories were respectively: 1) human 2) non-human 3) animate 4) date and time and 5) location. Accordingly, five different affix rules were defined for each of the above categories. आइतबार “Sunday” would now fall into the “date and time” category thus preventing it to get the द्वारा suffix applicable for the “human” category.

Furthermore, several new suffix rules have been defined in the affix file and correspondingly made association with applicable words in the dictionary file. Some of the suffixes for which new rules are made are the following – “ीकरण”, “ता”, “ीय”, “वाची”, “ात्मक”, “ीकृत”, “ीयता”, “िक”, “ान्तरण” .

The previous versions of the spell checker had not addressed the compound verbs effectively. In this version, we have identified seven different patterns of verbs basically covering compound verbs in Nepali and correspondingly devised rules for them.

4. Results

There has been a drastic improvement in this version of the Spell Checker. The Spell Checker now contains approximately 37,000 head words with

1,800 affix rules. With the enhancements introduced, the current version has the word coverage of approximately 6.2 million Nepali words. The suggestions that it provides also has been greatly refined. Major improvements can be seen in terms of postpositions and compound verbs this time. Random test of the Spell Checker in three different texts showed respectively 90% accuracy (43 words unhandled out of 450 words), 94% accuracy (25 words unhandled out of 400 words) and 89% accuracy (100 words unhandled out of 923 words). Currently, efforts are being made to fix the noted problems.

5. Conclusion

The current version of the Spell Checker with a substantial enhancements made is believed to have attained the industrial strength or robustness required for the target audience, i.e., publication houses, writers and journalists. Further testing and additional enhancements would be made to the Spell Checker in the days to come.

Acknowledgement

“This work was carried out with the aid of a grant from the International Development Research Centre, Ottawa, Canada administered through the centre for Research in Urdu Language Processing (CRLUP), National University of Computing and Emerging Sciences, Lahore, Pakistan (NUCES)”

6. References

[1] <http://www.sourceforge.net/projects/hunspell>

[2] Bal Krishna Bal et. al., "Nepali Spellchecker", *PAN Localization Working Papers 2004-2007*, Centre for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Lahore, Pakistan, pp. 316-318.

[3] Bal Krishna Bal et. al., "Nepali Spellchecker 1.1 and the Thesaurus, research and development", *PAN Localization Working Papers 2004-2007*, Centre for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Lahore, Pakistan, pp. 319-323.