Volume 4, Issue 6, June 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Hybrid Approach for Spell Checker and Grammar Checker for Punjabi

Jaspreet Kaur[*]
M.Tech
Department of Computer Science
Lovely Professional University, Phagwara, India

Kamaldeep Garg
Assistant Professor
Department of Computer Science
Lovely Professional University, Phagwara, India

Abstract*— Spell checker is a tool used for checking the spelling errors and also correcting those errors in the text or a document. Grammar checker is a program which is used to verify the grammatical errors in the text. Developing a spell checker and grammar checker for Indian languages such as Punjabi raises many new difficulties which are not in English, which makes the design of spell checker and grammar checker very difficult. The main difficulties faced are- there is no basic layout of Punjabi keyboard and also there is no approved format for Punjabi spellings. There are lots of differences in grammatical properties of Punjabi that makes it different from other languages. Punjabi is the world's 12th most widely spoken language. The very first requirement for developing any spell checker is to have dictionary of different words of that language which will work as lexicon. The paper aims to develop a system which is hybrid combination of spell checker and grammar checker for Punjabi. Firstly the system checks for spelling errors, then checks for grammatical errors in the text. When some input text is given to system, it is passed through spelling checker and then grammar checker. A comparative, efficient algorithm has been proposed which is hybrid combination of spell checker and grammar checker for Punjabi which saves time and cost.

***Keywords*— Error Detection, Error Correction, Spell Checker**

I. INTRODUCTION

Spell checker is a program or feature of program which is used to recognize the words which are misspelled and inform the user about those misspelled words. It depends on the feature of spell checker either to automatically correct the misspelled word or ask the user to choose the correct word from the suggestions which are available for that particular misspelled word. Spell checker may be an application which has capability of working on a large part of text or as an element of bigger application such as text editor, email, blog writing, keyword searching.

The main steps accomplished by spell checker are [4]:-

1. Receives the sentence or word as input.
2. Divide the sentence into words and also process the word to make it suitable for transferring it to the next step.
3. The word is searched in the lexicon.
4. If the word is present in the lexicon, next word is processed.
5. In case the word does not exist in the lexicon, then its closest matching words are searched for in the lexicon and then given to the user as suggestions.

Grammar checker is an application or an element of application that recognizes written text for grammatical errors and then corrects those errors. Almost all the grammar checkers are executed as an element of a bigger application, such as word document, email but they are also accessible as an application which may be standalone that can be prompted from within different types of programs that work with edible text. Natural language processing is mainly used for developing a grammar checker. An example of a software program that includes its own grammar checker is Microsoft Word [7].

II. PUNJABI LANGUAGE

Punjabi is most commonly written in the Gurmukhi script which is the most complete and accurate way to represent Punjabi sounds. Unlike Roman script, the Gurmukhi script follows a 'one sound-one symbol' principle. The Gurmukhi script has forty one letters including thirty eight consonants and three basic vowel sign bearers (Mataravahak). There are ten clear vowel signs and three auxiliary signs. The most striking characteristic of the Gurmukhi script, in comparison with Roman, is that, with the exception of five, all letters are joined by a line across the top. Like English and other European, Latin-based languages, it is written and read from left to right. However, there are neither capital letters in Gurmukhi nor articles such as 'a' and 'the'. Punjabi spellings are, for the most part, regular and relatively simple to learn. However, as is the case in English, Punjabi spellings are not fully standardized. Equivalent sounds which have been given in Romanized script are only approximate since the Gurmukhi script has many sounds unfamiliar to the English speaker which often may not be exactly represented by the Roman alphabet.

The Gurmukhi script, derived from the Sharada script and standardized by Guru Angad Dev in the 16th century was designed to write the Punjabi language. The meaning of "Gurmukhi" is literally "from the mouth of the Guru".

III. TECHNICAL ISSUES

A. *Issues with Spell Checker*

The first issue with spell checker is related to the error patterns which are produced by different types of text media such as typewriter and computer keyboard, OCR system, typesetting and machine printing, and handwriting. The error pattern of one media does not match with another. The error pattern problem of each media concerns the relative abundance of deletion, insertion, transposition and substitution error, run on and split word error, word length effect, positional bias, phonetic similarity effect etc. The knowledge about error pattern is necessary to model an efficient spell checker. Another issue is computerized dictionary which is related to the size of dictionary, dictionary file structure, dictionary partitioning, and word access techniques and so on [14].

.

B. *Issues with grammar checker*

The writing style of programs which were earlier available verified for wordy, common or the expressions which were misused in the text. This procedure was based simply on pattern matching. The actual grammar checking is very difficult. Natural languages do not have a specific syntax and grammar unlike computer languages that have specific syntax and grammar. Although it is feasible to write a grammar for a natural language which is completely formal, but there are many exceptions in real usage that while writing grammar checker, there is a minimal help by complete formal grammar. One of the most essential parts of a natural language grammar checker is a dictionary of all words in language, along with parts of speech of each word. The natural words can take many different parts of speech but it increases the difficulty of any grammar checker. A grammar checker verifies each word, then pre-processes that word, assign parts of speech tags to each word, then makes phrases and at last provides the user suggestions for the errors.

IV. MOST COMMON ERRORS IN PUNJABI

There are many errors that come across when human write Punjabi text and these errors generally belong to one of the following categories:

A. *Insertion error*

Insertion error occurs when one or more extra letters are inserted in the required word. For example: ਕਮਰ -> ਕਮਰਾ, ਸਤਾ->ਸੱਤਾ

In the above examples, ਕਮਰਾ and ਸੱਤਾ are also valid words but they are not required words. These types of errors can give rise to real word errors which means words are valid but not required. For example: ਉਠ->ਉੱਠ

B. *Deletion Error*

Deletion error occurs when one or more letter is removed from the required word. For example: ਉੱਠ -> ਉਠ, ਯੋਗਾ -> ਯੋਗ

ਉਠ and ਯੋਗ are valid words but they are not required.

These types of errors can also give rise to real word errors. For example: ਸ਼ੁੱਧੀ ->ਸ਼ੁੱਧ

C. *Substitution error*

Substitution error occurs when one or more letters are substituted by some another letter. ਆਮ -> ਆਪ, ਸਬਰ -> ਨਬਰ, ਸ਼ੀਸ਼ਾ -> ਸ਼ੀਸਾ.

In the above given examples, ਮ -> ਪ, ਸ -> ਨ, ਸ਼ -> ਸ are the various substitution pairs.

The main reasons for substitution errors are:

- Words which are generally used in many forms. For example: ਕੇਹਾ -> ਕਿਹਾ, ਮੇਹਨਤ -> ਮਿਹਨਤ
- Vowels which have similar sounds. For example: ਿ ->ੀ, ੁ - >ੂ, ੇ - >ੈ, ੋ ->ੌ
- Because of replacement of half characters. For example: ਰ ->੍ਰ, ਵ ->੍ਵ, ਹ ->੍ਹ

D. *Transposition Error*

Transposition error occurs when two adjacent letters are written in swapped way. For example: ਵੇਸਣ ->ਵਸੇਣ, ਸ਼ਾਮ -> ਸ਼ਮਾ

In the above explained examples, ੇ-> ਸ, ਾ -> ਮ

Transposition errors also give rise to real word errors (the word which are valid but not required.

E. *Run- on error*

Run-on error occurs when two or more valid words are erroneously typed side by side without a space in the middle of. For example: ਆਦਰ ਮਾਣ -> ਆਦਰਮਾਣ

ਦਾਦਾ ਜੀ -> ਦਾਦਾਜੀ, ਇਸ ਦਾ -> ਇਸਦਾ.

In the above explained examples, ਆਦਰ, ਮਾਣ, ਦਾਦਾ, ਜੀ, ਇਸ, ਦਾ are four different words. In some cases these words can also give rise to real word errors. For example: ਦਾਦਾ ਜੀ ਦਾਦਾਜੀ, ਇਸ ਦਾ, ਇਸਦਾ. Words ਦਾਦਾਜੀ, ਇਸਦਾ are two valid words.

F. Split word errors

Split word error is opposite of run-on error. These types of errors occur when there is some additional space is embedded between the parts of the word. It can be simply removed by deleting the additional space. For example: ਦੁਖੀ -> ਦੁ ਖੀ,ਨਿਰਭਰ -> ਨਿਰ ਭਰ. In some cases, split word errors can also give rise to real word errors. For example: ਇਸਦਾ -> ਇਸ ਦਾ, ਇਸਦੇ -> ਇਸ ਦੇ. Words ਇਸ, ਦਾ, ਇਸ, ਦੇ are four valid words.

G. Phonetically Similar Character Errors:

Phonetic errors are that type of errors in which the writer substitutes a phonetically correct but orthographically incorrect sequence of characters for the required word. It can be categorized into following types:

- Class 1: ਜ -> ਝ, ਬ -> ਭ, ਨ -> ਣ, ਗ-> ਘ
- Class 2: ਸ਼ -> ਸ, ਫ਼ -> ਫ, ਖ਼ -> ਖ, ਗ਼ -> ਗ, ਜ਼ -> ਜ, ਲ਼ -> ਲ
- Class 3: ਰ ->◌੍ਰ, ਵ ->◌੍ਵ, ਹ ->◌੍ਹ
- Class 4: ◌ੁ -> ◌ੂ, ◌ਂ -> ◌ੰ, ◌ੇ -> ◌ੈ,ਿ◌ -> ◌ੀ

V. MOTIVATION

Spell checker is very useful when user needs to write some kind of documents like applications, letters etc. in that case, user does not need to focus on the spelling errors. User need to gather data because spell checkers and grammar checkers automatically correct those mistakes done by the user. There are already available spell checkers available in Punjabi language. The scope of our study is to develop a system which is a hybrid combination of both spell checker and grammar checker. Spell checker alone can check only spelling errors but hybrid approach of spell checker and grammar checker will check both the spelling errors and grammatical errors of Punjabi. In order to develop this system, various existing algorithms of both are deeply studied and only this hybrid approach is proposed.

The research is focused on following objectives:

1. To analyse different types of errors in Punjabi text.
2. To create a corpus of many different Punjabi words this will act as dictionary
3. To develop the hybrid combination of spell checker and grammar checker.
4. To improve the accuracy of the resulting system.
5. To detect the spelling and grammatical errors of Punjabi text.
6. To change the erroneous word in the input text with correct suggestions.

VI. PROPOSED SCHEME

The paper discusses this important issue of energy optimization in hierarchically-clustered wireless sensor networks to minimize the total energy consumption required to collect data. The proposed system includes the routing protocols such as Leach and MAODV.

A. Existing System

As spell checkers and grammar checkers helps us to find both the spelling and grammatical errors in written words. Both spell checker and grammar checker are basic necessity of writing text in every language. Though appreciable work has been done in English and associated languages, the Indian language scenario present a relatively more complex and uphill task. Punjabi is the world's 12th most widely spoken language [4]. The available spell checkers for Punjabi language are 'AKHAR', 'SUDHAAR', and 'RAFTAAR' and there is one grammar checker available for Punjabi language.

AKHAR is bilingual spell checker for English and Punjabi. AKHAR can automatically detect the language and invokes the respective spell checker. It provides the facility to select the words from alternatives list by using Alt keys only. But it is paid software that is not available free for its use to everybody [14].

Designing a spell checker and grammar checker for Punjabi poses many challenges as there is very much difference in Punjabi language when it is compared with other languages. There are algorithms available for spell checker but in those spell checkers, only non-word errors are mainly considered. There are mainly two types of errors: non word and real word errors. A non-word error is a word that is not a valid word or that does not exists in the dictionary. A real word error is a word which exists in the lexicon but it is not intended in the sentence. The research work is mainly focused on combining both the spell checker and grammar checker and also minimizing both these types of errors. This hybrid approach of spell checker and grammar checker for Punjabi aims at minimizing the time as well as cost. This system will be very helpful for the users.

B. Proposed Method

Present work is mainly focused on developing a hybrid approach of spell checker and grammar checker for Punjabi. This approach combines two algorithms of spell checker and grammar checker which will be very helpful in saving the time. While writing the text, it will be first passed through spell checker and if errors come, spell checker will firstly correct those errors and then grammatical errors are checked.

Basically in this work, firstly statement will be firstly passed through a spell checker and after correcting the spellings, it will be passed through grammar checker. The main steps of spell checker: Dictionary Creation, Pre-processing, Error Detection, displaying the suggestions and then replacing the error with most suitable suggestion. The steps included in grammar checker are: Morphological analyser, parts of speech tagger, phrase chunker and then displaying the correct output. While implementing the spell checker, the erroneous word is firstly checked in the standard dictionary and the dictionary or lexicon created by the user. If the word is not found in the lexicon, it will be treated as an error. After that, suggestion list will be generated and then list will be sorted displaying the suggestions. Then final output will be displayed after correcting the errors. After this, it will be passed through grammar checker which will firstly assign POS (Parts of speech) tags, then it will disambiguate the information and then it will make various phrases. The accuracy of this system is expected to be greater than existing algorithms.

C. Methodology

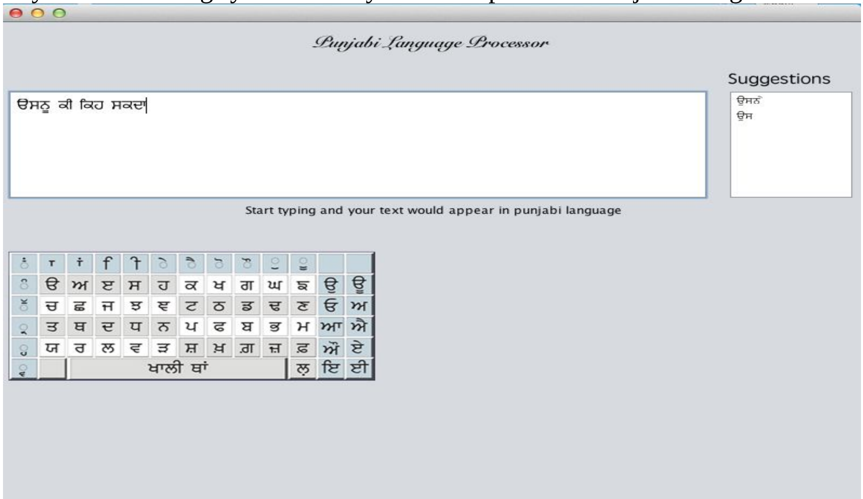
The procedure to be followed in the research is:

i. Create a dictionary (lexicon) of commonly occurring words which will be referred by the spell checker for comparing the words and then after that for providing the suggestions to the user about incorrect word.
ii. Differentiate words on the basis of the spaces. Spaces can be considered as the character between two words. As soon as the user presses the space bar, it will provide suggestion list.
iii. When a word is found. Run the search to find the word in the dictionary (this is done in a thread so that the typing of the user is not affected)
iv. If the word is found then no action happens and if it is not found we run an algorithm that tries to find out what the user had to write (i.e. the correct word).
v. After generating the suggestion list, the system sorts the suggestion list so that the user may get the suggestion in the useful format.
vi. Finding the correct word by matching the first two letters of the word and trying to find all words that are there.
vii. Then the user selects the word from the suggestion list and then replaces it in the input text.
viii. At this step, user will get the text in which there is no spelling error.
ix. As soon as user presses "," or "|", the grammar checking module starts.
x. For grammar checking, the grammatically incorrect sentence is passed through pre-processing module.
xi. This text is passed through morphological analyser which uses full form lexicon to assign each word it's all possible part-of-speech information [7].
xii. Then the text along with POS tags is moved to POS tagger which attempts to disambiguate the information.
xiii. Then the POS tagged text is passed on to phrase chunker which builds phrases.
xiv. Then the syntax/agreements checks are performed based on the grammatical information.
xv. At last, any discrepancy that is found, is reported to the user and the corrections along with the detailed error information will be provided to the user

The proposed methodology is needed to be implemented in a simulation environment. The system is implemented in java using Netbeans7 platform.

VII. RESULTS AND DISCUSSION

Evaluating a system is as important as to develop it. So here we are going to evaluate our Hybrid system. In this we are checking the accuracy of the resulting system. The system is implemented in java using netbeans7 platform.



Punjabi Language Processor

Suggestions

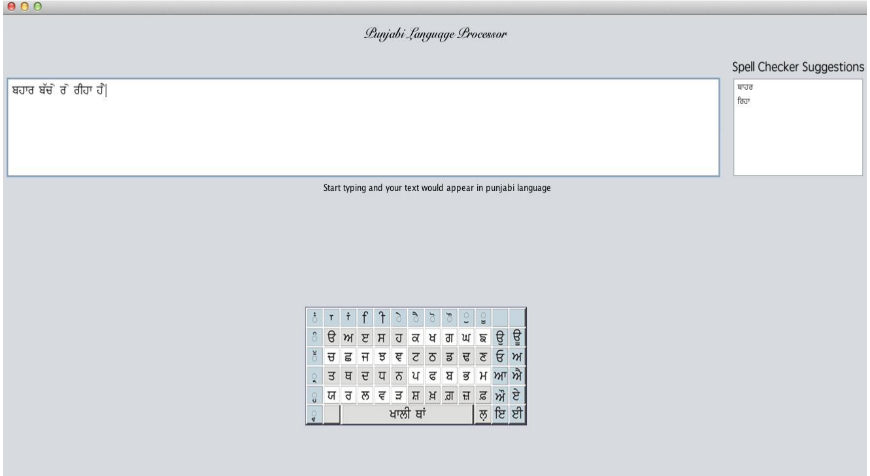
ਉਸਨੂ ਕੀ ਕਿਹ ਸਕਦਾ

ਉਸਨੰ
ਉਸ

Start typing and your text would appear in punjabi language

Fig. 1 user providing input

In this user provides the sentence as input and it will provide suggestions according to the errors in the given input. It will check the presence of the word in the dictionary and provide the suggestions by matching first two characters and then same thing for the remaining characters.

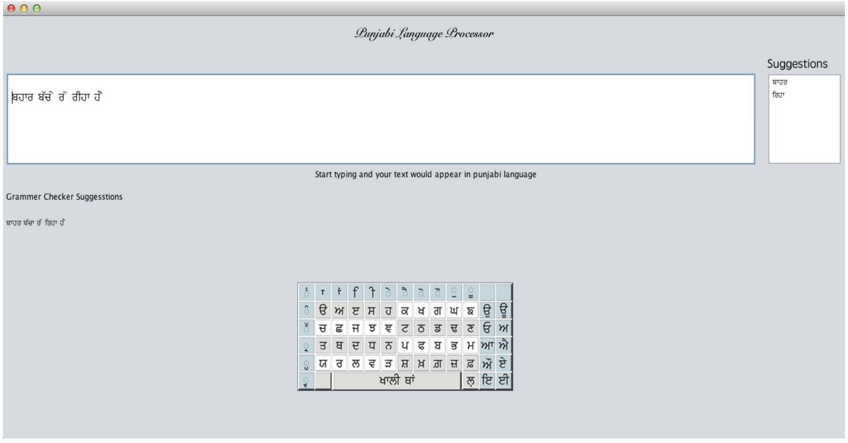

Punjabi Language Processor
Spell Checker Suggestions
Start typing and your text would appear in punjabi language
ਖਾਲੀ ਥਾਂ

ਬਹਾਰ ਬੱਚੇ ਰੋ ਰੀਹਾ ਹੈ

ਬਾਹਰ
ਰਿਹਾ

Fig. 2 User output

In this sentence, there are two spelling errors: ਬਹਾਰ and ਰੀਹਾ. All the words are matched with the database which contains the valid Punjabi words. Since these are incorrect words so the correct words will be suggested in the Spell Checker Suggestions column. The closest suggestions that are found from database are ਬਾਹਰ, ਰਿਹਾ Therefore these words are displayed as output of spell checker module.


Punjabi Language Processor
Suggestions
Start typing and your text would appear in punjabi language
Grammer Checker Suggesstions
ਖਾਲੀ ਥਾਂ

ਬਹਾਰ ਬੱਚੇ ਰੋ ਰੀਹਾ ਹੈ

ਬਾਹਰ
ਰਿਹਾ

ਬਾਹਰ ਬੱਚਾ ਰੋ ਰਿਹਾ ਹੈ

Fig. 3 User output

This is grammatically incorrect sentence. In this modifier or noun ਬੱਚੇ is not in conformance with the verb ਰਿਹਾ ਹੈ. Root word of ਬੱਚੇ is ਬੱਚਾ. It should be in singular instead of plural. So the input sentence will be written as:ਬੱਚਾ ਰੋ ਰਿਹਾ ਹੈ

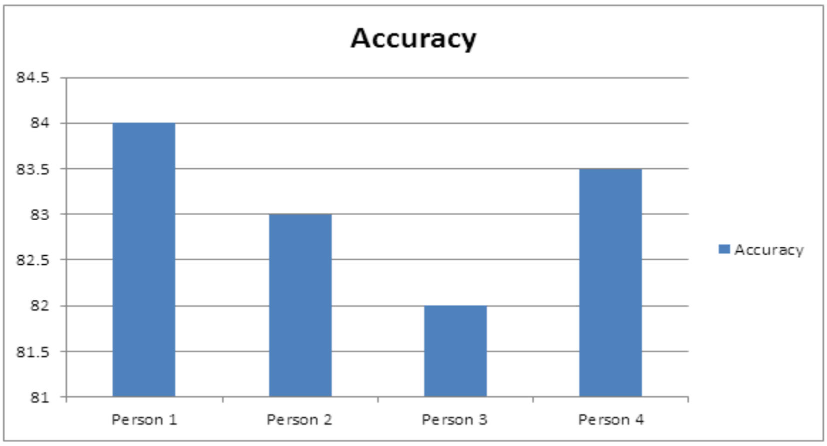


Fig. 4 Accuracy graph

The average accuracy of our resulting system is 83.5 %.

VIII. CONCLUSION

The intent of the paper is to propose a system which is hybrid combination for spell checker and grammar checker for Punjabi language. The proposed system firstly checks the spelling errors and after correcting spelling errors, then after this it checks the grammatical errors. The final output is a text which is both spelling error and grammatical error free. The average accuracy of the resulting system is 83.5 %. For future scope, this system can be enhanced for the complex sentences. The accuracy of the system can also be improved. This system can be used for other languages also.

.

ACKNOWLEDGMENT

The paper has been written with the kind assistance, guidance and active support of my department who have helped me in this work. I would like to thank all the individuals whose encouragement and support has made the completion of this work possible.

REFERENCES

[1] Aarti Tyal, Dharam Veer Sharma "Punjabi Thesaurus-a tool for Natural Language Processing", Research Journal of Computer Systems Engineering- An International Journal, Vol 02, Issue 02, June, 2011.

[2] Amit Sharma, Pulkit Jain (2013) "Hindi Spell Checker", Indian Institute of Technology Kanpur, April 17, 2013.

[3] Dr. Baldev Singh Baddn (2001), National Punjabi Kosh, National Publishers Ltd., New Delhi.

[4] Gurpreet Singh Lehal (2007),"Design and Implementation of Punjabi Spell Checker", International Journal of Systematic cybernetics and informatics, pp.70-75.

[5] Gurpreet Singh Lehal & Meenu Bhagat (2007), "Spelling Error Pattern Analysis of Punjabi Typed Text", In Proceedings of International Symposum on Machine Translation, NLP and TSS, pp.128-141, 2007.

[6] Lata Bopche, Gauri Dhopavakar (2012) "Rule Based Grammar Checking System For Hindi", Journal of Information Systems and Communication ISSN: 0976-8742 & E-ISSN: 0976-8750, Volume 3, Issue 1, pp- 45-47.

[7] Mandeep Singh Gill, Gurpreet Singh Lehal,"A Grammar Checking System for Punjabi", Department of Computer Science, Punjabi University, Patiala.

[8] Md. Jahangir Alam, Naushad UzZaman and Mumit Khan "N-gram based Statistical Grammar Checker for Bangla and English", Center for Research on Bangla Language Processing, BRAC University, Dhaka, Bangladesh.

[9] Naushad UzZaman, Mumit Khan "A Comprehensive Bangla Spelling Checker", Center for Research on Bangla Language Processing BRAC University, Bangladesh.

[10] Neha Gupta, Pratistha Mathur (2012) "Spell Checking Techniques in NLP: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 12, December 2012, pp. 217-221.

[11] Rada Mihalcea, Hugo Liu and Henry Lieberman, "NLP (Natural Language Processing) for NLP (Natural Language Programming)", Computer Science Department, University of North Texas, Media Arts and Sciences, Massachusetts Institute of Technology

[12] Ritika Mishra, Navjot Kaur (2013), "Design and Implementation of Online Punjabi Spell Checker Based on Dynamic Programming", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3.

[13] Ronan Collobert, Jason Weston "A Unified Architecture for Natural Language Processing: Deep Neural Networks With Multitask Learning", NEC Labs America, 4 Independence Way, Princeton, NJ 08540 USA

[14] Rupinderdeep Kaur, Prateek Bhatia (2010) "Design and Implementation of SUDHAAR-Punjabi Spell Checker", International Journal of Information and Telecommunication Technology, Vol. 1, Issue 1,2010, (0976-5972).