A Punjabi Grammar Checker

Mandeep Singh Gill
Department of Computer Science
Punjabi University
Patiala -147002, India
msgill_in@yahoo.com
Tel.: +91-9888165971

Gurpreet Singh Lehal
Department of Computer Science
Punjabi University
Patiala -147002, India
gslehal@yahoo.com
Tel.: +91-175-3046171

Shiv Sharma Joshi
Department of Anthropological Linguistics & Punjabi Lexicography
Punjabi University
Patiala -147002, India
Tel.: +91-175-3046292

Abstract

This article provides description about the grammar checking software developed for detecting the grammatical errors in Punjabi texts and providing suggestions wherever appropriate to rectify those errors. This system utilizes a full-form lexicon for morphology analysis and rule-based systems for part of speech tagging and phrase chunking. The system supported by a set of carefully devised error detection rules can detect and suggest rectifications for a number of grammatical errors, resulting from lack of agreement, order of words in various phrases etc., in literary style Punjabi texts.

1 Introduction

Grammar checking is one of the most widely used tools within natural language engineering applications. Most of the word processing systems available in the market incorporate spelling, grammar, and style-checking systems for English and other foreign languages, one such rule-based grammar checking system for English is discussed in (Naber, 2003). However, when it comes to the smaller languages, specifically the Indian languages, most of such advanced tools have been lacking. Spell checking has been addressed for most of the Indian languages but still grammar and style checking systems are lacking. In this article a grammar checking system for Punjabi, a member of the Modern Indo-Aryan family of languages, is provided. The grammar checker uses a rule-based system to detect grammatical errors in the text and if possible generates suggestions to correct those errors.

To the best of our knowledge the grammar checking provided here will be the first such system for Indian languages. There is n-gram based grammar checking system for Bangla (Alam et al, 2006). The authors admit its accuracy is very low and there is no description about whether the system provides any suggestions to correct errors or not. It is mentioned that it was tested to identify correct sentences from the set of sentences provided as input but nothing is mentioned as far as correcting those errors is concerned. However, the system that we discuss here for Punjabi detects errors and suggests corrections as well. In doing so, provides enough information for the user to understand the error reason and supports the suggestions provided, if any.

2 System Overview

The input Punjabi text is given to the preprocessing system that performs tokenization and detects any phrases etc. After that morphological analysis is performed, this returns possible tags for all the words in the given text, based on the full-form lexicon that it is using. Then a rule-based part of speech tagger is engaged to disambiguate the tags based on the context information. After that, the text is grouped into various phrases accordingly to the pre-defined phrase chunking rules. In the final phase, rules to check for various grammatical errors internal to phrases and agreement on the sentence level, are applied. If any error is found in a sentence then based on the context information corrections are suggested (generated) for that.

For the purpose of morphological analysis we have divided the Punjabi words into 22 word classes like noun, adjective (inflected and uninflected), pronoun (personal, demonstrative, reflexive, interrogative, relative, and indefinite), verb (main verb, operator verb, and auxiliary verb), cardinals, ordinals, adverb, postposition, conjunction, interjection etc., depending on the grammatical information required for the words of these word classes. The information that is in the database depends upon the word class, like for noun and inflected adjective, it is gender, number, and case. For personal pronouns, person is also required. For main verbs gender, number, person, tense, phase, transitivity etc. is required. As mentioned earlier the lexicon of this morphological analyzer is full form based i.e. all the word forms of all the commonly used Punjabi words are kept in the lexicon along with their root and other grammatical information.

For part of speech tagging, we have devised a tag set keeping into mind all the grammatical categories that can be helpful for agreement checking. At present, there are more than 600 tags in the tag set. In addition to this, some word-specific tags are also there. The tag set is very user friendly and while choosing tag names existing tag sets for English and other such languages were taken into consideration, like NNMSD – masculine, singular, and direct case noun, PNPMPOF – masculine, plural, oblique case, and first person personal pronoun. The approach followed for part of speech tagging is rule-based, as there is no tagged corpus for Punjabi available at present. As the text we are processing may have grammatical agreement errors, so the part of speech tagging rules are devised considering this. The rules are applied in sequential order with each rule having an attached priority to control its order in this sequence.

For phrase chunking, again a rule-based approach was selected mainly due to the similar reasons as for part of speech tagging. The tag set that is being used for phrase chunking includes tags like NPD – noun phrase in direct case, NPNE – noun phrase followed by ਨੇ ne etc. The rules for phrase chunking also take into account the potential errors in the text, like lack of agreement in words of a potential phrase. However, as would be expected there is no way to take the misplaced words of a phrase into account, like if words of a phrase are separated (having some other phrase in between) then that cannot be taken as a single phrase, even though this may be a potential error.

In the last phase i.e. grammar checking, there are again manually designed error detection rules to detect potential errors in the text and provide corrections to resolve those errors. For example, rule to check modifier and noun agreement, will go through all the noun phrases in a sentence to check if the modifiers of those sentences agree with their respective head words (noun/pronoun) in terms of gender, number, and case or not. For this matching, the grammatical information from the tags of those words is used. In simple terms, it will compare the grammatical information (gender, number, and case) of modifier with the headword (noun/pronoun) and displays an error message if some grammatical information fails to match. To resolve this error, the grammar checking module will use morphological generator, to generate the correct form (based on headword's gender, number, and case) for that modifier from its root word.

For example, consider the grammatically incorrect sentence ਸੋਹਣੇ ਲੜਕਾ ਜਾਂਦਾ ਹੈ sohne larka janda hai 'handsome boy goes'. In this sentence in the noun phrase, ਸੋਹਣੇ ਲੜਕਾ sohne larka 'handsome boy', the modifier ਸੋਹਣੇ sohne 'handsome' (root word – ਸੋਹਣਾ sohna 'handsome'), with masculine gender, plural number, and direct case, is not in accordance with the gender, number, case of its head word. It should be in singular number instead of plural. The grammar checking module will detect this as an error as 'number' for modifier and headword is not same, then it will use morphological generator to generate the 'singular number form' from its root word, which is same as root form i.e. ਸੋਹਣਾ sohna 'handsome' (masculine gender, singular number, and direct case). So, the input sentence will be corrected as ਸੋਹਣਾ ਲੜਕਾ ਜਾਂਦਾ ਹੈ sohna larka janda hai 'handsome boy goes'.

The error detection rules in grammar checking module are again applied in sequential order with priority field to control the sequence. This is done to resolve phrase level errors before going on to the clause level errors, and then to sentence level agreement errors.

3 Grammar Errors

At present, this grammar checking system for Punjabi detects and provides corrections for following grammatical errors, based on the study of Punjabi grammar related texts (Chander, 1964; Gill and Gleason, 1986; Puar, 1990):

Modifier and noun agreement

The modifier of a noun must agree with the noun in terms of gender, number, and case. Modifiers of a noun include adjectives, pronouns, cardinals, ordinals, some forms of verbs etc.

Subject and verb agreement

In Punjabi text, the verb must agree with the subject of the sentence in terms of gender, number, and person. There are some special forms of verbs like transitive past tense verbs, which need some specific postpositions with their subject, like the use of ਨੇ ne with transitive verbs in perfect form etc.

Noun and adjective (in attributive form) agreement

This is different from 'modifier and noun agreement' as described above in the sense that adjective is not preceding noun but can be virtually anywhere in the sentence, usually preceding verb phrase acting as a complement for it. It must still agree with the noun for which it is used in that sentence.

Order of the modifiers of a noun in noun phrase

If a noun has more than one modifier, then those modifiers should be in a certain order such that phrase modifiers precede single word modifiers but pronouns and numerals precede all other.

Order of the words in a verb phrase

There are certain future tense forms of Punjabi verbs that should occur towards the end of verb phrase without any auxiliary. In addition, if negative and emphatic particles are used in a verb phrase then the latter must precede the former.

ਦਾ da postposition and following noun phrase agreement

All the forms of ਦਾ da postposition must agree in terms of gender, number, and case with the following noun phrase that it is connecting with the preceding noun phrase.

Some other options covered include noun phrase must be in oblique form before a postposition, all the noun phrases joined by connectives must have same case, main verb should be in root form if preceding ਕੇ ke etc.

4 Sample Input and Output

This section provides some sample Punjabi sentences that were given as input to the Punjabi grammar checking system along with the output generated by this system.

Sentence 1

Shows the grammatical errors related to 'Modifier and noun agreement' and 'Order of the modifiers of a noun in noun phrase'. In this sentence noun is ਲੜਕਾ larka 'boy' and its modifiers are ਸੋਹਣੀ ਇੱਕ ਭੱਜੀ ਜਾਂਦਾ sohni ek bhajji janda 'handsome one running'.

Input: ਸੋਹਣੀ ਇੱਕ ਭੱਜੀ ਜਾਂਦਾ ਲੜਕਾ ਆਇਆ
Input1: sohni ek bhajji janda larka aaeya
Input2: Handsome one running boy came
Output: ਇੱਕ ਭੱਜਿਆ ਜਾਂਦਾ ਸੋਹਣਾ ਲੜਕਾ ਆਇਆ
Output1: ek bhajjia janda sohna larka aaeya
Output2: One running handsome boy came

Sentence 2

Covers the grammatical error related to 'Subject and verb agreement'. Subject is ਬਾਰਸ਼ barish 'rain' and verb phrase is ਹੋ ਰਿਹਾ ਹਨ ho riha han 'is raining'.

Input: ਬਾਹਰ ਬਾਰਸ਼ ਹੋ ਰਿਹਾ ਹਨ
Input1: bahr barish ho riha han
Input2: It is raining outside
Output: ਬਾਹਰ ਬਾਰਸ਼ ਹੋ ਰਹੀ ਹੈ
Output1: bahr barish ho rahi hai
Output2: It is raining outside

Sentence 3

For grammatical errors related to 'ਦਾ da postposition and following noun phrase agreement' and 'Noun phrase in oblique form before a post position'. Noun phrase preceding ਦੀ dee

(possessive marker) is ਛੋਟਾ ਬੱਚਾ chota baccha ‘small boy’ and following one is ਨਾਮ naam ‘name’.

Input: ਛੋਟਾ ਬੱਚਾ ਦੀ ਨਾਮ ਰਾਮ ਹੈ

Input1: chota baccha dee naam raam hai
Input2: Small boy’s name is Ram
Output: ਛੋਟੇ ਬੱਚੇ ਦਾ ਨਾਮ ਰਾਮ ਹੈ

Ouput1: chote bacche da naam raam hai
Ouput2: Small boy’s name is Ram

Sentence 4

Highlights the grammatical errors related to ‘Subject and verb agreement’ and ‘Order of the words in a verb phrase’. Subject in this sentence is ਲੜਕੀ larki ‘girl’ and verb phrase is ਨਹੀਂ ਜਾ ਹੀ ਰਿਹਾ ਸੀ nahi ja hee riha see ‘was not going’.

Input: ਲੜਕੀ ਸਕੂਲ ਨਹੀਂ ਜਾ ਹੀ ਰਿਹਾ ਸੀ

Input1: larkee school nahi ja hee riha see
Input2: The girl was not going to school
Output: ਲੜਕੀ ਸਕੂਲ ਜਾ ਹੀ ਨਹੀਂ ਰਹੀ ਸੀ

Ouput1: larkee school ja he nahi rahi see
Output2: The girl was not going to school

Sentence 5

For grammatical error related to ‘Subject and verb agreement’. Subject here is ਰਾਮ raam ‘Ram’ and verb phrase is ਖਾਧਾ khadha ‘ate’, which is transitive and in perfect phase.

Input: ਰਾਮ ਫ਼ਲ ਖਾਧਾ

Input1: raam phal khadha
Input2: Ram ate fruit
Output: ਰਾਮ ਨੇ ਫ਼ਲ ਖਾਧਾ

Output1: raam ne phal khadha
Output2: Ram ate fruit

Legend:

- **Input** and **Output** specifies the input Punjabi sentence in Gurmukhi script and the output produced by this grammar checking system in Gurmukhi script, respectively.
- **Input1/Output1** specifies the Romanized version of the **input/output**.
- **Input2/Output2** specifies the English gloss for the **input/output**.

5 System Features

The system is designed in Microsoft Visual C# 2005 using Microsoft .NET Framework 2.0. The entire database of this tool is in XML files with the Punjabi text in Unicode format. Some of the significant features of this grammar checking system are:

Rules can be turned on and off individually

Being a rule-based system all the rules provided in section 3 can be turned on and off individually without requiring any changes in the system. The rules are kept in a separate XML file, not hard coded into the system. To turn on/off a rule, changes can be made to that XML file directly or it can be done through the options provided within the system.

Error and Suggestions information

The system is able to provide enough reasons in support of every error that it detects. With a meaningful description of the rule, it provides the grammatical categories that failed to match if there is an error and provides the desired correct value for those grammatical categories, with suggestions. However, the information about grammatical categories may not be much meaningful to an ordinary user but if someone is learning Punjabi as a foreign/second language then information about correct grammatical categories according to the context can be helpful. Wherever possible system also specifies both the words, for which matching was performed, making it more clear that what is wrong and with respect to what, as shown in Figure 1, it shows that which was the head word and which word failed to match with it.

The suggestions produced by the Punjabi Grammar Checker for the following grammatically incorrect sentence to correct the first incorrect word ਰਹੀਆਂ rahian ‘-ing plural’ are ਰਿਹਾ riha ‘-ing singular’ and ਰਹੀ rahi ‘-ing singular’:

ਮੈਂ ਖੇਡ ਰਹੀਆਂ ਹਨ

main khed rahian han ‘I are playing’

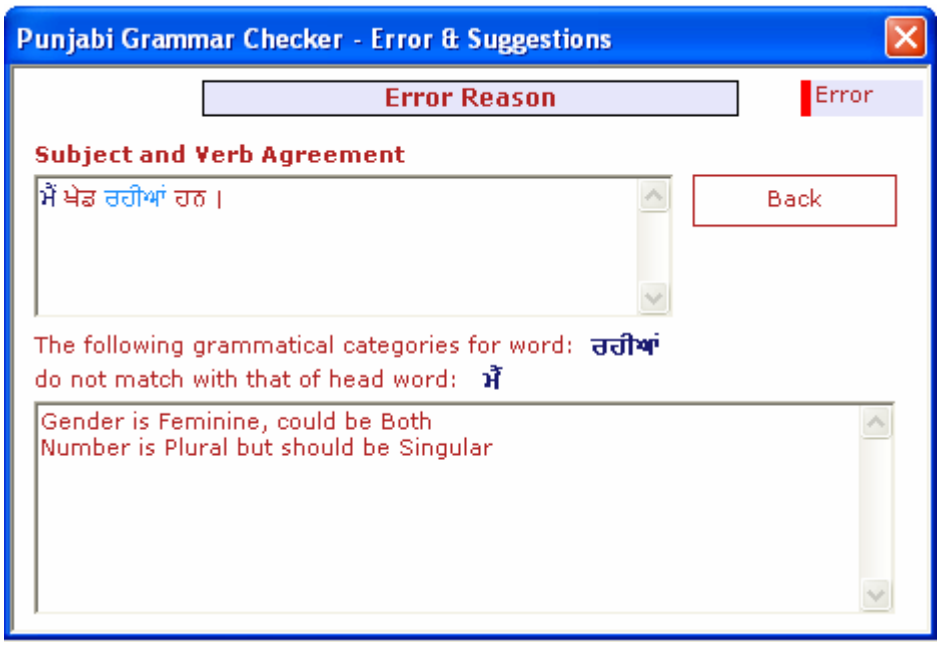

Figure 1. Punjabi Grammar Checker – Error Reason

Figure 1 shows the grammatical categories that failed to match for the subject ਮੈਂ main 'I' and part of the verb phrase ਰਹੀਆਂ rahian '-ing plural'. It provides the values for the grammatical categories that failed to match for the incorrect word along with the desired values for correction.

6 System Scope

The system is designed to work on the literary style Punjabi text with SOV (Subject Object Verb) sentence structure. At present, it works properly on simple or kernel sentences. It can detect any agreement errors in compound or complex sentences also. However, there may be some false alarms in such sentences. The sentences in which word order is shuffled for emphasis has not been considered, along with the sentences in which intonation alone is used for emphasis. Due to emphatic intonation, the meaning or word class of a word may be changed in a sentence e.g., ਤੇ te 'and' is usually a connective but if emphasized it can be used as an emphatic particle. However, this is hard to detect from the written form of the text and thus has not been considered. However, if some emphatic particles like ਹੀ he ਈ ee ਵੀ ve etc., are used directly in a sentence to show emphasis then that is given due consideration.

7 Hardware and Software Requirements

The system needs hardware and software as would be expected from a typical word processing application. A Unicode compatible Windows XP based PC with 512 MB of RAM, 1 GB of hard disk space and Microsoft .NET Framework 2.0 installed, would be sufficient.

References

Duni Chander. 1964. *Punjabi Bhasha da Viakaran (Punjabi).* Punjab University Publication Bureau, Chandigarh, India.

Daniel Naber. 2003. *A Rule-Based Style and Grammar Checker*. Diplomarbeit Technische Fakultät, Universität Bielefeld, Germany. (Available at: http://www.danielnaber.de/languagetool/download/style_and_grammar_checker.pdf (1/10/2007))

Harjeet S. Gill and Henry A. Gleason, Jr. 1986. *A Reference Grammar of Punjabi*. Publication Bureau, Punjabi University, Patiala, India.

Joginder S. Puar. 1990. *The Punjabi verb form and function*. Publication Bureau, Punjabi University, Patiala, India.

Md. Jahangir Alam, Naushad UzZaman, and Mumit Khan. 2006. N-gram based Statistical Grammar Checker for Bangla and English. In *Proc. of ninth International Conference on Computer and Information Technology (ICCIT 2006)*, Dhaka, Bangladesh.