A Grammar Checking System for Punjabi

Mandeep Singh Gill
Department of Computer Science
Punjabi University
Patiala – 147002, India
msgill_in@yahoo.com

Gurpreet Singh Lehal
Department of Computer Science
Punjabi University
Patiala – 147002, India
gslehal@yahoo.com

Abstract

This article provides description about the grammar checking system developed for detecting various grammatical errors in Punjabi texts. This system utilizes a full-form lexicon for morphological analysis, and applies rule-based approaches for part-of-speech tagging and phrase chunking. The system follows a novel approach of performing agreement checks at phrase and clause levels using the grammatical information exhibited by POS tags in the form of feature value pairs. The system can detect and suggest rectifications for a number of grammatical errors, resulting from the lack of agreement, order of words in various phrases etc., in literary style Punjabi texts. To the best of our knowledge, this grammar checking system is the first such system reported for Indian languages.

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 widely used languages. Naber (2003) discussed one such rule-based grammar checking system for English. However, when it comes to the smaller languages, specifically the Indian languages, most of such advanced tools have been lacking. Although, spell checking has been addressed for most of the Indian languages, still grammar and style checking systems are lacking. In this article, a grammar checking system for Punjabi has been provided. Punjabi is a member of the Modern Indo-Aryan family of languages.

There is an n-gram based grammar checking system for Bangla (Alam et al., 2006). However, the authors admit that its accuracy is very low and there is no description about whether the system provides any suggestions to correct errors or not. However, the system that we discuss here for Punjabi detects errors and suggests corrections as well. While doing so, it provides enough information for the user to understand the error reason and the suggestions provided, if any.

2 Purpose

The purpose of the system is to find various grammatical mistakes in the formal texts written in Punjabi language. While detecting grammatical mistakes, the focus is on keeping the false alarms to minimum. For every detected error, system provides enough information for the user to understand why the error is being marked. It also provides suggestions, if possible, to rectify those errors.

3 Potential Applications

This system as a whole and its subsystems will find numerous applications in natural language processing of Punjabi. Following are some of the application areas of this system as a whole or its subsystems:

- It can be used with various information processing systems for Punjabi, where the input needs to be corrected grammatically before processing.

Coling 2008: Companion volume – Posters and Demonstrations, pages 149–152
Manchester, August 2008

- Parts of this system like morphological analyzer, morphological generator, part-of-speech tagger, phrase chunker etc., will find use in almost every natural language processing application like machine translation, text to speech synthesis, and search engines etc., for Punjabi.
- This system as a whole can be used as a post editor for a number of applications for Punjabi like machine translation, optical character recognition etc., where the output needs to be corrected grammatically before providing the end results.
- Second language learners of Punjabi can use this system as a writing aid to learn grammatical categories operating in Punjabi sentences, and thus improve their writings by learning from their mistakes.
- In the word processing field, this system can be used for checking essays, formal reports, and letters written in Punjabi.

4 System Design & Implementation

The design of this grammar checking is provided below in figure 1. A sketchy idea of this proposed design is provided below in terms of how the input text is processed to find potential grammatical errors.

For grammar checking, the input text is first given to a preprocessor, which breaks the input text into sentences and then into words. It also performs filtering, i.e. marks any phrases, fixed expressions etc. in the text. Then the tokenized text is passed on to a morphological analyzer, which uses a full form lexicon to assign each word its all possible part-of-speech (POS) information (i.e. POS tags). Then the text along with the POS tags moves on to a POS tagger, which attempts to disambiguate the information using hand-written disambiguation rules. Then this POS tagged text is passed on to a phrase chunker, which builds phrases using hand-written phrase chunking rules targeted at the POS tag information. Phrase chunker also marks clause boundaries and headwords in noun phrases and clauses. Then in the last stage, syntax/agreement checks are performed based on the grammatical information (exhibited by POS tags) at the phrase level and then at the clause level, using the marked headwords. Any discrepancy found is reported to the user along with the suggested corrections and detailed error information.

All the sub activities of this grammar checking system are fully automated and have been designed exclusively from scratch as part of this work. No such sub system was available for Punjabi for our use. All the sub activities have been implemented in Microsoft Visual C# 2005 and the databases are in XML format with Punjabi text in Unicode.

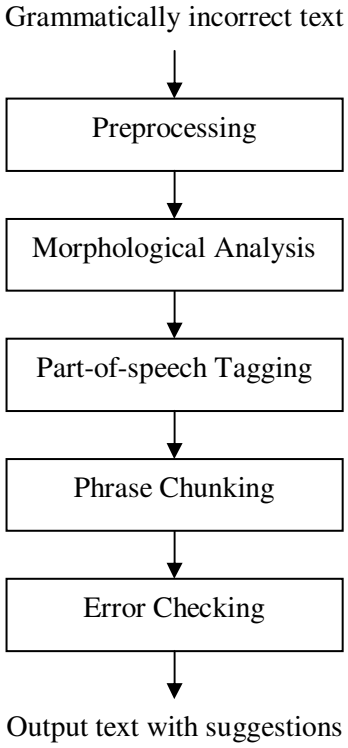

Figure 1. Punjabi Grammar Checking System Design

For the purpose of morphological analysis, we have divided the Punjabi words into 22 word classes 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. As mentioned earlier, all the word forms of 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. 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. The part-of-speech tagging rules take into account the potential grammatical agreement errors.

For phrase chunking, again a rule-based approach was selected. 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 take into account the potential errors in the text, like lack of agreement in words of a potential phrase.

In the last phase i.e. error checking, there are 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 headwords (noun/pronoun) in terms of gender, number, and case or not. For this matching, the grammatical information from the tags of modifiers and headwords 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 error 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, and case of its headword ਲੜਕਾ *larka* 'boy'. 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'.

5 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. **Input/Output** specify the input Punjabi sentence and the output produced by this grammar checking system respectively.

Sentence 1

This sentence 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

This sentence 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

6 Testing & Evaluation

The evaluation results for our morphological analyzer shows that it provides correct analysis for 87.64% words. This evaluation was performed on a corpus of 8 million Punjabi words. The part-of-speech tagger reports an accuracy of 80.29% when applied on a randomly selected corpus of 25,000 words. This accuracy improves to 88.86% if we exclude unknown

words from evaluation results, the reason being the absence of an unknown word guesser in our part-of-speech tagger. The phrase chunker reports average precision of 81.18%, recall of 85.07%, and F-measure of 83.07%. These results include evaluation performed on 100 sentences for noun, adjective, and verb phrases. On randomly selected 1,000 sentences, this grammar checking system reports precision of 76.79%, recall of 87.08%, and F-measure of 81.61%.

The grammatical errors covered by this system includes – modifier and noun agreement, subject/object and verb agreement, order of modifiers in a noun phrase, order of words in a verb phrase, use of contractions etc. In its present state, the system may generate some false alarms for complex and compound sentences. We will work to reduce these false alarms in the future.

Comparison with existing systems

Our system covers a different class of errors and results for grammar checking systems for English, Swedish etc. are reported for different error sets, with only some errors covered being common. Some of the systems that are to some extent close to our system in terms of errors covered are provided here for comparison. A grammar checker for German (Schmidt-Wigger, 1998) using pattern matching rules reports 81% precision and 57%. A system for Korean (Young-Soog, 1998) reports 99.05% precision and 95.98% recall. Another system for German (Fliedner, 2002) reports precision and recall of 67% for only noun phrase agreement. A grammar checker for Bangla (Alam et al., 2006) reports accuracy of 53.7% using manual POS tagging and 38% for automated POS tagging. When compared with these systems, 76.79% precision and 87.08% recall of our grammar checker seams reasonably good.

7 Conclusions

This article presented design and implementation details of the grammar checking system for Punjabi. This grammar checking system is capable of detecting various grammatical errors in formal Punjabi texts. To the best of our knowledge, this is the first such system for Punjabi and other Indian languages. We hope that this research work will attempt to narrow down the gap that exists between Punjabi and other natural languages in the natural language processing field. We are confident that this research work will motivate future researchers in developing various advanced resources for Punjabi. This article presented a novel approach for performing grammar checking using phrase and clause level information coupled with grammatical information (POS information) in the form of feature values. This approach can be applied for languages that lack advanced resources like full parser, and pattern-matching approaches are not competent enough to detect different agreement errors.

The web-based version of this grammar checking is available for free use along with three other resources for the Punjabi language – morphological analyzer, part-of-speech tagger, and phrase chunker. Morphological analyzer is also available as free download for non-commercial use.

References

Alam, Md. Jahangir, 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.

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

Fliedner, Gerhard. 2002. A System for Checking NP Agreements in German Texts. In *Proceedings of the ACL Student Research Workshop*, pages 12-17, Philadelphia, US.

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

Naber, Daniel. 2003. *A Rule-Based Style and Grammar Checker*. Diplomarbeit Technische Fakultät, Universität Bielefeld, Germany.

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

Schmidt-Wigger, Anje. 1998. Grammar and Style Checking for German. In *Proceedings of the Second International Workshop on Control Language Applications (CLAW-1998)*, pages 76-86, Pittsburgh, PA.

Young-Soog, Chae. 1998. Improvement of Korean Proofreading System Using Corpus and Collocation Rules. In *Proceedings of the 12th Pacific Asia Conference on Language, Information and Computation*, pages 328-333, National University of Singapore, Singapore.