IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-8 E-ISSN: 2347-2693

Detection and Correction of Style Errors Present in Punjabi Sentences

S. K. Sharma

Dept. of Computer Science and Applications, DAV University, Jalandhar, India

*Corresponding Author: sanju3916@rediffmail.com

Available online at: www.ijcseonline.org

Received: 07/Jul/2017, Revised: 20/Jul/2017, Accepted: 15/Aug/2017, Published: 30/Aug/2017

Abstract-- Detection and correction of style errors in a language plays an important role in development of language related resources like Machine Translation, Grammar checking, Natural Language Interfaces etc. Style errors may include various types of errors like Missing sentence ender, Wrong sentence ender, Error due to missing comma, Repeated/duplicate word Error, error due to missing conjunction etc.. Though considerable work has been done in the area for English and related languages, but the Indian Language scenario presents a relatively more complex and uphill task. In this paper, author has presented an algorithm for detection of various style errors present in Punjabi language. Author tested his algorithm using three different kinds of data sets and it is observed that the algorithm performs better for wrong sentence ender and duplicate word type error as compare to missing conjunction type style error.

Keywords-- Style error, grammar checker, Punjabi Language processing, NLP.

I. INTRODUCTION

Detection and correction of style error is important part of language proofing. These errors may generate incorrect interpretation of the sentence. As per (Naber, 2003) Style errors are results of using uncommon words and complicated sentence structures. This makes a text more difficult to understand, which is seldom desired. These cases are thus considered style errors. Unlike grammar errors, it heavily depends on the situation and text type which cases should be classified as a style error. Basically four types of errors are included in the style error type. These four types are: Missing sentence ender, Wrong sentence ender, Error due to missing comma, Repeated/duplicate word Error and error due to missing conjunction.

A. Missing sentence ender:

This type of error occurs due to missing punctuation mark at the end of the sentence. Consider the following example:

Incorrect Sentence

Punjabi: ਤੁਸੀ ਕਿਹੜੇ ਸਕੂਲ ਜਾਣਾ ਹੈ

Transliterated: (tusī kihṛē sakūl jāṇā hai)
Translated: In which school you have to go

In above sentence there is no sentence ender. Above sentence is an interrogative sentence, but instead of using a question mark (?) as the sentence ender, no sentence ender has been used. The correct sentence should be:

Correct Sentence

Punjabi: ਤੁਸੀ ਕਿਹੜੇ ਸਕੂਲ ਜਾਣਾ ਹੈ?

Transliterated: (tusī kihṛē sakūl jāṇā hai)?
Translated: In which school you have to go?

B. Wrong sentence ender:

This type of error occurs due to use of an incorrect punctuation mark or missing punctuation mark in the sentence. Consider the following example:

Incorrect Sentence

Punjabi: ਤੁਸੀ ਕਿਹੜੇ ਸਕੂਲ ਜਾਣਾ ਹੈ।

Transliterated: (tusī kihṛē sakūl jāṇā hai).
Translated: In which school you have to go.

Above sentence is an interrogative sentence, but instead of using a question mark (?) as the sentence ender, an affirmative sentence ender (। *'dandi'*) has been used. Therefore, it is incorrect use of punctuation mark (। *'dandi'*). The correct sentence should be:

Correct Sentence

Punjabi: ਤੁਸੀ ਕਿਹੜੇ ਸਕੂਲ ਜਾਣਾ ਹੈ?

Transliterated: (tusī kihṛē sakūl jāṇā hai)?
Translated: In which school you have to go?

C. Repeated/duplicate word Error

Sometime, while writing a sentence, a word is typed twice. This results in unstructured sentence and it becomes difficult

to understand the meaning of the sentence. Consider the following example:

Incorrect Sentence

Punjabi: ਜਦੋਂ ਉਸ ਦੀ ਜਾਗ ਖੁੱਲ੍ਹੀ ਖੁੱਲ੍ਹੀ ਚੋਰੀ ਹੋ ਚੁੱਕੀ ਸੀ
duplicate word

Transliterated: (jadōṃ us dī jāg khullhī khullhī cōrī hō cukkī sī)

In the above sentence, ਖੁੱਲ੍ਹੀ (khullhī) is a duplicate word as it has been typed twice. One of these duplicate words should be removed from the sentence. The correct sentence should be:

Correct Sentence

Punjabi: ਜਦੋਂ ਉਸ ਦੀ ਜਾਗ ਖੁੱਲ੍ਹੀ, ਚੋਰੀ ਹੋ ਚੁੱਕੀ ਸੀ

Transliterated: (jadōṃ us dī jāg khullhī cōrī hō cukkī sī)

D. Error due to missing comma or conjunction

Compound sentences are composed of independent clauses separated by conjunctions. These conjunctions include some punctuation marks like comma (,) or they may be words belonging to coordinate conjunctions word class like ਅਤੇ, ਤਾਂ etc. If this conjunction is missing in the sentence, then two clauses will merge into single clause and it becomes difficult to process the compound sentence. Consider the following example:

Incorrect Sentence

Punjabi: ਜਦੋਂ ਉਸ ਦੀ ਜਾਗ ਖੁੱਲ੍ਹੀ ਚੋਰੀ ਹੋ ਚੁੱਕੀ ਸੀ

Transliterated: (jadōṃ us dī jāg khullhī cōrī hō cukkī sī)
Translated: When she woke up the robbery had already happened.

In the above sentence, there are two clauses; one is adverb clause i.e. ਜਦੋਂ ਉਸ ਦੀ ਜਾਗ ਖੁੱਲ੍ਹੀ (jadōṃ us dī jāg khullhī) and second is independent clause i.e. ਚੋਰੀ ਹੋ ਚੁੱਕੀ ਸੀ (cōrī hō cukkī sī). The two clauses should be separated by comma i.e. there should be comma (,) after the adverb clause "ਜਦੋਂ ਉਸ ਦੀ ਜਾਗ ਖੁੱਲ੍ਹੀ" (jadōṃ us dī jāg khullhī). Therefore, this sentence has a missing punctuation mark. The correct sentence should be:

Correct Sentence

Punjabi: ਜਦੋਂ ਉਸ ਦੀ ਜਾਗ ਖੁੱਲ੍ਹੀ, ਚੋਰੀ ਹੋ ਚੁੱਕੀ ਸੀ

Transliterated: (jadōṃ us dī jāg khullhī, cōrī hō cukkī sī)
Translated: When she woke up, the robbery had already happened.

II. WORK DONE RELATED TO ERROR DETECTION AND CORRECTION

Kukich[6] has discussed the various techniques for automatically detection and correction of misspellings and the various factors affecting the spelling errors patterns of words in English. Chaudhuri and Kundu [7]have done a detailed analysis on error pattern generated by Bangla text patterns and made a reversed word dictionary and phonetically similar word grouping based spellchecker for Bangla text. Church and Gale [8] have done a Probability scoring for spelling correction. Damerau [9] worked on a technique for computer detection and correction of spelling errors in English language. Morris and Cherry [10] devised an alternative technique for using trigram frequency statistics to detect errors. Pollock and Zamora [11] aimed at discovering probabilistic tendencies, such as which letters and position within a word are most frequently involved in errors, with the intent of devising a similarity key based technique. Yannakoudakis and Fawthrop [12-13] sought a general characterization of misspelling behavior. Wagner [14] was the first one to introduce the notion of applying dynamic programming techniques to the spelling correction problem to increase computational efficiency. A "reverse" minimum edit distance technique was used by Gorin [15] in the DEC-10 spelling corrector and by Durham et al. [16] in their command language corrector. Kernighan et al [17] and Church and Gale [18] also used a reverse technique to generate candidates for their probabilistic spelling corrector. This is the first time that a detailed error analysis for Punjabi is being carried out. For this purpose we have collected about 20000 misspelled words generated by typists, both novice and experienced as well as students learning Punjabi typing. We have done analysis of six main categories of errors. These errors are discussed in detail in following sections.

III. ALGORITHM USED

Step1: Scan the sentence from left to right. If it contains auxiliary verb then it will be marked with the sentence ender.

Step2: If the sentence starts with the interrogative word, then sentence should end with the question mark.

Step3: If a sentence starts with exclamatory word then the sentence should end with the exclamatory mark.

Step4: If within a sentence, there are more than one auxiliary verbs, then there should be comma or conjunction after the first auxiliary verb.

Various types of grammatical mistakes in an independent clause are handled by researcher's system are listed in table 1. First column of the table represents the type number of the error, second column represents the name of the grammatical mistake and third column shows the example containing incorrect and correct sentences related with the corresponding error shown in second column.

Table 1: Various types of grammatical mistakes due to style error in an independent clause

Sr. No	Type of error	Example	Remarks
1.	Missing sentence ender	ਖੇਡ ਦੇ ਮੈਦਾਨ ਵਿਚ ਖਿਡਾਰੀ ਸਹਿਯੋਗ ਨਾਲ ਖੇਡਦੇ ਹਨ (khēḍ dē maidān vic khiḍārī sahiyōg nāl khēḍdē han)	The sentence ender is missing.
2.	Wrong sentence ender	ਕੀ ਤੁਸੀਂ ਆਪਣੀ ਪੜ੍ਹਾਈ ਪੂਰੀ ਕਰ ਲਈ ਹੈ।(kī tusīṃ āpṇī paṛhāī pūrī kar laī hai.)	It is interrogative sentence and should be ended with a question mark (?)
3.	Error due to missing comma	ਇਕ ਦੋਸਤ ਹੀ ਕਾਫੀ ਹੁੰਦਾ ਹੈ ਦੋ ਦੋਸਤਾਂ ਜਿਹੀ ਕੋਈ ਰੀਸ ਨਹੀਂ, ਤਿੰਨ ਦੋਸਤ ਕਰਮਾਂ ਵਾਲਿਆਂ ਦੇ ਹੁੰਦੇ ਹਨ, ਚਾਰ ਦੋਸਤ ਸੰਭਵ ਨਹੀਂ। (ik dōsat hī kāphī hundā hai dō dōsatāṃ jihī kōī rīs nahīṃ, tinn dōsat karmāṃ vāliāṃ dē hundē han, cār dōsat sambhav nahīṃ .)	Comma is missing after the first independent clause i.e. ਇਕ ਦੋਸਤ ਹੀ ਕਾਫੀ ਹੁੰਦਾ ਹੈ (ik dōsat hī kāphī hundā hai)
4.	Repeated/duplicate word Error.	ਉਹ ਸਕੂਲ ਸਕੂਲ ਜਾ ਕੇ ਪੜ੍ਹਨ ਲੱਗ ਪਿਆ। (uh sakūl sakūl jā kē paṛhan lagg piā.)	Word ਸਕੂਲ (sakūl) has been typed twice.
5.	Error due to missing conjunction	ਮੁੰਡੇ ਨੇ ਰੋਟੀ ਖਾਧੀ ਘਰ ਚਲਿਆ ਗਿਆ।(muṇḍē nē rōṭī khādī ghar caliā giā.)	Conjunction is missing after the first independent clause i.e. ਮੁੰਡੇ ਨੇ ਰੋਟੀ ਖਾਧੀ (muṇḍē nē rōṭī khādī)

IV. RESULT AND DISCUSSION

For testing various types of style errors present in the Punjabi text, three types of test data have been designed. These are:

- Dummy test data with all possible errors incorporated manually.
- Real data from test/exam sheets of primary school students learning Punjabi as second language.
- The output of Hindi to Punjabi (H2P) machine translation system.

A. Dummy test data:

This data has been developed manually by inducing the errors in the sentences. This dummy test data has been developed by incorporating each type of error in the sentences. For instance, for creating a dummy test data to detect noun adjective agreement error, sentences having disagreement in number, gender and case between noun and adjective have been created. Two sets have been created i.e. one, by inducing single error per sentence and second, by inducing multiple errors per sentence. Table 2 shows total number of sentences with total number of words used for testing.

*B. Real Data***:**

In case of real data, incorrect sentences have been collected from workbooks of students who are learning Punjabi as second language. Then, these sentences have been categorized into different error types. Those sentences have also been included in the test data which do not contain any grammatical mistakes. These sentences have been used to identify the false alarm raised by the system.

*C. Hindi To Punjabi (H2p) Machine Translation System Output***:**

One of the applications of this style error checking system is to improve the performance of Hindi to Punjabi machine translation system by correcting its output. Therefore, the output of H2P machine translation system has been included in the test data. Again, this test data reveals the performance of our style error checking system. Like that of real data, this test data also contains grammatically correct sentences so that any false alarm raised by the system could be detected. Table 2 shows the detail of the number of sentences with number of words and table 3 shows amount of data taken for testing the system.

Table 2: Types of test data

Test data type	Number of sentences	Number of words
Dummy test data	5030	25309
Real test data	1000	6490
H2P Machine translation output	2000	12430

Table 3.1: Dummy Data taken for testing the system

Sr.No	Sub-category of style error	No. of sentences from dummy test data	Corrected by system	%age accuracy
1	Missing comma	85	83	97.6
2	Missing conjunction	76	65	85.5
3	Wrong sentence ender	150	148	98.6
4	Duplicate word	231	231	100

Table 3.2: Real Data taken for testing the system

Sr. No	Sub-category of style error	No. of sentences from real data	Corrected by system	%age Accuracy
1	Missing comma	45	38	84.4
2	Missing conjunction	34	23	67.6
3	Wrong sentence ender	23	23	100
4	Duplicate word	12	12	100

Table 3.3: H2P Machine Translation Data taken for testing the system

Sr. No	Sub-category of style error	No. of sentences from Hindi to Punjabi Machine translation system	Corrected by system	%age Accuracy
1	Missing comma	111	82	73.8
2	Missing conjunction	85	62	72.9
3	Wrong sentence ender	57	57	100
4	Duplicate word	30	30	100

Table 3.1, 3.2, 3.3 shows the result obtained by the style error correction system. Results show that the system almost 100% accuracy in case of wrong sentence ender and duplicate word type style errors, but its performance decreases for missing comma and missing conjunction. This reduction in performance is due to the fact that in complex Punjabi sentences, there is no auxiliary verb between dependent and independent clauses and hence it becomes difficult for the computer system to identify the location to insert comma or conjunction. In future, work can be done on this.

REFERENCES

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

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

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

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

[5]. 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.

[6]. K. Kukich (1992) "*Techniques for automatically correcting words in text*", ACM Computing Surveys. 24(4): 377-439.

[7]. P. Kundu and B.B. Chaudhuri, "*Error Pattern in Bangla Text*". International Journal of Dravidian Linguistics, Vol.28, Issue.2, pp.49-88, 1999.

[8]. K.W. Church and W.A. Gale, "*Probability scoring for spelling correction*", Statistical Computing, Vol.1, Issue.1, pp.93-103, 1991.

[9]. F.J. Damerau, "*A technique for computer detection and correction of spelling errors*", Commun. ACM. Vol.7, Issue.3, pp.171-176, 1964.

[10]. Morris, Robert and Cherry, Lorinda L, "*Computer detection of typographical errors",* IEEE Trans Professional Communication, vol. PC-18, no.1, pp54-64, March 1975.

[11]. Pollock, J. J., Zamora, A, "*Collection and characterization of spelling errors in scientific and scholarly text",* J. Amer. Soc. Inf. Sci. Vol.34, Issue.1, pp.51-58, 1983.

[12]. Yannakoudakis, E. J., Fawthrop, D., "*An intelligent spelling corrector*", Inf. Process. Manage. 19, 12, 101-108, 1983.

[13]. Yannakoudakis, E.J. and Fawthrop, D, "*An intelligent spelling error corrector*", Information Processing and Management, vol.19, no.2, pp.101-108, 1983.

[14]. Wagner, Robert A. and Fischer, Michael J, "*The string-to-string correction problem",* Journal of the A.C.M., vol.21, no.1, pp168-173, January 1974.

[15]. R.E. Gorin, "*SPELL: A spelling checking and correction program*", Online documentation for the DEC-10 computer, 1971.

[16]. Durham, I, Lamb, D.A, and Saxe, J.B, "*Spelling correction in user interfaces*", Communications of the A.C.M., vol.26, no.10, pp764-773, October 1983.

[17]. M.D. Kernighan, K.W. Church, and W.A. Gale., *"A spelling correction program based on a noisy channel model*", In Proceedings of the Thirteenth International Conference on Computational Linguistics, pp.205-210, 1990.

[18]. Gale and Church, William A. Gale and Kenneth W. Church., "*A program for aligning sentences in bilingual corpora*", In Proceedings of the 29th Meeting of the ACL, pages 177-184. Association for Computational Linguistics, 1991.

[19]. Roger Mitton, "*Spelling checkers, spelling correctors and the misspellings of poor spellers*", Information Processing and Management: an International Journal, Vol..23 Issue.5, pp.495-505, Sept. 1987.

Author Profile

Dr. S.K. Sharma pursed Bachelor of Engineering from SLIET (Sant Longowal Institute of Engineering and Technology), Longowal in 2000 and Master of Technology in Computer Science and Engineering from Punjabi University Patiala in year 2008. He completed his Ph.D.from Punjabi University Patiala in 2016 and currently working as Assistant Professor in Department of Computer Science and Applications at DAV University Jalandhar since 2013. He has published more than 24 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online. His main research work focuses on Syntactic Analysis, POS tagging Algorithms, Data Mining and Computational linguistics. He has 14 years of teaching experience and 7 years of Research Experience.

