Various Techniques Used For Grammar Checking

Blossom Manchanda
Assistant Professor
Department of CSE
GNDEC Ludhiana

Vijay Anant Athavale
Director
Gulzar Group of Institutions
Khanna

Sanjeev kumar Sharma
Assistant professor
DAV University
Jalandhar

ABSTRACT

Grammar checker is one of proofing tool used for syntactic analysis of the text. Various techniques are used for development of grammar checker. These techniques includes rule based technique, statistical based technique and syntax based technique. In this research article, all these three techniques have been discussed. Both advantages and disadvantages of these techniques have also been discussed at the end.

Keywords

Grammar checker, rule based, statistical technique, syntax based technique

1. INTRODUCTION

Grammar checker of a language is a system that detects various grammatical errors in a given text based on the grammar of that particular language, and reports those errors to the user along with a list of helpful suggestions to rectify those errors. It is also expected to provide the error details including the error reason to explain that why a particular error is being dubbed a grammatical error. A grammar checker not only detects grammatical errors, but also suggests possible corrections. Grammar checkers are also part of the fundamental tools line word processor needed for NLP in any language. Till date best supported language in NLP is English. A lot of work has gone into developing sophisticated systems that have gone into widespread use, such as automatic translators, grammar checker and spell checkers.

2. SYNTAX BASED TECHNIQUE

In this approach, full parsing is performed on the given text.After parsing, each sentence is assigned a tree structure based on the underlying language's grammar. If complete parsing is not succeeded, then the text is considered as incorrect. Therefore, in order to reduce the number of false alarm, the parser should be as complete as possible.The main advantage of this approach is that, if the grammar specified is complete, i.e. covers all the possible syntactic rules of the language, then the grammar checker will detect all the incorrect sentences, irrespective of the nature of the error. However, due to the ambiguities in natural languages, it is not possible to list explicitly all the syntax rules of a natural language. Moreover, the parser may return more than one parse tree even for correct sentences. Using this approach one can only detect incorrect sentences. However, to tell the user what the problem is, some extra rules are required that parse ill-formed sentences, this technique is known as constraint relaxation. If a sentence can only be parsed using such an extra rule, then the sentence is termed as incorrect and accordingly rule explanation and suggestions can be provided. A list of syntax based grammar checker developed for different languages and researcher is provided in table 1.

Table 1: List of syntax grammar checker developed using rule based technique

Sr. No.	Language	Author	Year
1	Korean	Young-Soog	1998
2	Danish	Paggio	(1999, 2000)
3.	French	Vandeventer	2001
4.	Urdu	Kabir et. Al.	2002

3. STATISTICS BASED TECHNIQUE

In this approach, an annotated corpus is used to generate a list of part-of-speech (POS) tag. From these generated sequences, some sequences will be very common (for example *adverbverb*), others will probably not occur at all (for example *verb adjective*). Commonly occurring sequences will be considered correct in other texts also and uncommon sequences will lead to errors. One major pre-requisite for applying this approach is the availability of substantial amount of POS-annotated corpus. In addition, the POS tags used must reflect the grammatical properties required for checking agreement in the

underlying language. For instance, for grammar checking of the Hindi text we cannot use the Punjabi corpus annotated with the POS tags designed for the English language. Alam et al. (2006) reported poor accuracy while applying English tagset for corpus-based grammar checking of Bangla. Now if we consider the Hindi language sentence then, adjective and noun must agree in terms of gender, number, and case, in a noun phrase. Therefore, POS tags for both the adjective and the noun word classes must exhibit these grammatical categories, if used for tagging the corpus for use in grammar checking of Hindi. Because if we use just a single *adjective* tag for tagging adjectives and a single *noun* tag for tagging nouns, then the grammar checker based on this corpus's statistics may term all the adjective and noun pairs as correct. Irrespective of the fact that some of those pairs may not be in agreement with each other, for some or all of the above-mentioned grammatical categories.Major problem with this approach like other statistics-based systems is that the results are difficult to interpret. If there is a false alarm then the user will wonder why the input is being termed incorrect, as there are no specific error messages. The developer will also need access to the training corpus to understand the system's opinion. Another problem with this approach is that someone has to set a threshold to separate the 'uncommon but correct constructs' from 'uncommon and incorrect ones'. A list of statistics based grammar checker developed for different languages and researcher is provided in table 2.

Table 2: List of statistical grammar checker developed using rule based technique

Sr. No.	Language	Author	Year
1	English	Park et al.	1997
2	French	Tschichold et al.,	1997
3	English	Powers	1997
4	Brazilian Portuguese	Martins et al.	1998
5	Swedish	Arppe	1999
6	Bangla and English	Alam et al.	2006
7	Swedish language	Sjöbergh	2006
8	Persian language	Ehsan and Faili	2010
9	Amharic language	Temesgen and Assabie	2012
10	A Language Independent Statistical Grammar (LISG) checking system	Verena Henrich and Timo Reuter	2009

4. RULE BASED TECHNIQUE

In this approach, a set of predefined rules in the form of error patterns are matched against a text, which has at least been POS tagged. Text is erroneous if a match is found for one of those patterns. Patterns can be based directly on words, their POS tags, or even chunk tags. This approach is similar to the statistics-based approach, but all the rules are developed manually.A rule-based system unlike syntax-based system will never be complete. As it is almost impossible to foresee all the grammatical inconsistencies, so there will always be some errors it does not find, even if numerous error rules are present. However, leaving some errors undetected is still better than incomplete parser raising annoying false alarms.This approach has few advantages over other approaches, as each rule can be turned on and off individually and system can offer detailed error messages along with helpful comments, even explaining grammar rules. This approach allows incremental building up of the system, starting with just one rule and then extending it rule by rule. A list of rule based grammar checker developed for different languages and researcher is provided in table 3.

Table 3: List of rule based grammar checker developed using rule based technique

Sr. No.	Language	Author	Year
1	Dutch	by Vosse	1992
2	Czech and Bulgarian	by Kuboň and Plátek	1994
3	Swedish	by Hein	1998
4	French, German, and Spanish.	Helfrich and Music	2000

5	French	Vandeventer	2001
6	Swedish	Carlberger et al.	2002, 2004
7	English	Naber	2003
8	Brazilian Portuguese	Kinoshita et al.	2006
9	Nepali	Bal and Shrestha	2007
10	Persian	Ehsan and Faili	2010
11	Chinese	Jiang et al.	2011
12	Malay	Kasbon et al.	2011

5. COMPARISON BETWEEN DIFFERENT APPROACHES

All these three approached have their own advantages and disadvantages. Advantaged and disadvantages of all three approaches are summarized in table 4.

Table 4: Advantages and disadvantages of various techniques

Technique	Advantages	Disadvantages
Rule based technique	▪ It is easy to incorporate domain knowledge into linguistic knowledge ▪ easy to understand ▪ user can easily extend the rules for handling new error types ▪ Rules can be built incrementally by starting with just one rule and then extending it ▪ Each rule of a rule-based system can be easily configured ▪ provides detailed analysis of the learner's writing using linguistic knowledge and provides reasonable feedback ▪ the linguistic knowledge acquired for one natural language processing system may be reused to build knowledge required for a similar task in another system	▪ Complexity of the grammar increases exponentially as we try to solve different types of errors. ▪ need a lot of manual effort ▪ increases cognitive load on the human analyst and also increases the degree of ambiguity in the grammar ▪ requires complete grammar rules to cover all types of sentence constructions
Statistics based technique	▪ When the training set and the test set are similar, then ML approach provides good results. ▪ No need of deep knowledge of grammar. ▪ Language independent system can be developed.	▪ Data sparseness poses a problem for ML ▪ Most of the time, ML based system does not provide necessary comments on errors. ▪ Users are usually surprised when system predicts a correct sentence as wrong. ▪ Results of ML based systems are difficult to interpret. ▪ Sometimes debugging the reasons of system's failure becomes very

		complicated because the results are generated by aggregating probabilities and frequencies. ▪ Some ML based systems rely on threshold values which are usually estimated heuristically. Threshold may vary depending on the domain of text where the system is trained or tested.
Syntax based technique	▪ Both grammatical and ungrammatical sentences can be parsed using constraint relaxation ▪ The errors in an ungrammatical sentence can be easily identified based on the constraints which are relaxed during parsing of the sentence	▪ constraint relaxation technique is not well suited for parsing sentences with missing or extra words ▪ Failure of parsing does not always reliably ensure that the input sentence is ungrammatical because the insufficient coverage of grammatical rules may also be a cause of unsuccessful parsing ▪ robust parsers with sufficient linguistic rules are not available ▪ rule-based parsers suffer from the curse of natural language ambiguities which unnecessarily produce more than one parse tree even for the correct input sentence

6. CONCLUSION

From above discussion, it can be concluded that all techniques have their own advantages and disadvantages. Therefore selection of technique to be used for developing a grammar checker depends upon type of language and existing resources of the language. If annotated corpus is available for a language then statistical techniques can be applied. Similarly if all possible rules of the grammar of a language can be easily developed then rule based technique can be used. If parser is available then syntax based technique can be preferred.

REFERENCES

[1]. Mallikarjun, B, Yoonus, M. Sinha, Samar & A. Vadivel,"Indian Languages and Part-ofSpeech Annotation. Mysore", Linguistic Data Consortium for Indian Language, pp. 22-25. ISBN-81-7342-197-8, 2010.
[2]. Alam, M. J., Uzzaman, N., & Khan, M. (2006). N-gram based Statistical Grammar Checker for Bangla and English. Ninth International Conference on Computer and Information Technology (ICCIT 2006), 3–6.
[3]. B. K. Bal, B. Pandey, L. Khatiwada, P. Rupakheti, and M. P. Pustakalaya, "Nepali grammar checker," PAN/L10n/PhaseII/Reports by Madan PuraskarPustakalaya, Lalitpur, Nepal, pp. 1–5, 2008.
[4]. LataBopche, Gauri Dhopavkar, and ManaliKshirsagar, "Grammar Checking System Using Rule Based Morphological Process for an Indian Language", Global Trends in Information Systems and Software Applications, 4th International Conference, ObCom 2011 Vellore, TN, India, December 9-11, 2011.
[5]. F. R. Bustamante, "GramCheck: a grammar and style checker," COLING '96 Proc. 16th Conf. Comput. Linguist., vol. 1, pp. 175–181, 1996.
[6]. D. Tesfaye, "A rule-based Afan Oromo Grammar Checker," IJACSA - Int. J. Adv. Comput. Sci. Appl., vol. 2, no. 8, pp. 126–130, 2011.
[7]. J. Carlberger, R. Domeij, V. Kann, and O. Knutsson, "A Swedish grammar checker," Citeseer, 2000.
[8]. N. Ehsan and H. Faili, "Statistical Machine Translation as a Grammar Checker for Persian Language," Sixth Int. Multi-Conference Comput. Glob. Inf. Technol., no. c, pp. 20–26, 2011.
[9]. Y. Jiang, T. Wang, T. Lin, F. Wang, W. Cheng, X. Liu, C. Wang, and W. Zhang, "A rule based Chinese spelling and grammar detection system utility," Syst. Sci. Eng. (ICSSE), 2012 Int. Conf., no. 1, pp. 437–440, 2012.
[10]. M. Gill, G. Lehal, and S. Joshi, "A Punjabi Grammar Checker," Acl.Eldoc.Ub.Rug.Nl, pp. 940–944.
[11]. H. Kabir, S. Nayyer, J. Zaman, and S. Hussain, "Two Pass Parsing Implementation for an Urdu Grammar Checker."
[12]. J. Kaur, "Hybrid Approach for Spell Checker and Grammar Checker for Punjabi," vol. 4, no. 6, pp. 62–67, 2014.

[13]. N. Ehsan and H. Faili, "Towards grammar checker development for Persian language," Proc. 6th Int. Conf. Nat. Lang. Process. Knowl. Eng. NLP-KE 2010, 2010.
[14].]K. F. Shaalan, "Arabic GramCheck: A grammar checker for Arabic," Softw. - Pract. Exp., vol. 35, no. 7, pp. 643–665, 2005.
[15]. J. Kinoshita, C. Eduardo, and D. De Menezes, "CoGrOO: a Brazilian-Portuguese Grammar Checker based on the CETENFOLHA Corpus," October, pp. 2190–2193, 2003.
[16]. D. Deksne and R. Skadiņš , "CFG Based Grammar Checker for Latvian," Proc. 18th Nord. Conf. Comput. Linguist. NODALIDA 2011, p. 275 278, 2011.