LEXICO-SEMANTIC ANALYSIS OF ESSAYS IN HINDI LANGUAGE

Seema Mahato
Research Scholar,
Dr. C.V. Raman University,
BILASPUR (C.G.), INDIA,
seema_mahato@yahoo.co.in

Dr. (Mrs.) Ani Thomas
Professor, Dept. of Computer Applications,
Bhilai Institute of Technology,
DURG (C.G.), INDIA,
tpthomas22@yahoo.com

ABSTRACT

Large number of researchers consider essay as a tool to judge learning outcomes and intellectual capabilities and to assess the organized and integrated thoughts. Due to increase in the number of universities students and distance and ubiquitous e-learning approaches, the intention of using Computer-based Assessment Systems has rise rapidly in a decade. Manual grading of students' essays requires significant amount of time and hard work and also an expensive activity for educational institutions and need a practical solution to this task. The automated essay grading or evaluation system is solution to such need. So now-a-days, most of the online competitive and universities exams are trying to evaluate the human written essays by examiners / teachers as well as by machines like automated essay grading system. Such system has to significantly focus on vocabulary and text syntax, and text semantics. The research paper focus on the existing automated essay grading systems, their functional technologies and proposes a methodology to overcome the issues related to them while evaluating such as grammatical and semantic error as well as influence of local and regional languages in Hindi essays.

CCS Concepts

•Computing methodologies →Artificial intelligence → Natural language processing → Lexical semantics

Keywords

Automated Essay Grading, AEG, NLP, Text Processing, Essay Evaluation, Semantic Attributes

1. INTRODUCTION

Mostly the essay grading systems available on the market are used for grading essays written in pure English or pure European languages. In India, we have almost 21 recognized languages and 27 local languages and influence of these languages can be easily seen in Hindi essays. For examples, in the text “दिवाली के दिन लक्ष्मी के आगमन के विश्वास के साथ लोग अपने घरो के आँगन मे रंगोली या ***आलपोना*** से सजाते है” and “राखी के पुण्य पर्व पर घर मे किसी के निधन से यह त्योहार ***खोटा*** हो जाता है”, the influence of regional or local language can be clearly identified. Presently Computer-based Assessment Systems (CbAS) are rare whose primary focus is to provide automatic grading and evaluation of Hindi essays. Automated Essay Grading (AEG) Systems, also known as Automated Writing Evaluation Systems or Automated Essay Assessors. The aim of such system is to score student’s essays on a specific topic and give feedback to the student on deficiencies in his/her essay. These systems lower down the burden of the evaluator as unable to give personalized attention to the student’s needs. Such systems provide human capability of reading and writing and also time-to-time feedback to the writers/students which help them to improve their writing skill.

2. STATE-OF-ART

Currently available systems to the automated assessment are Project Essay Grade (PEG), Intelligent Essay Assessor (IEA), Educational Testing service I (ETS I), Electronic Essay Rater (E-Rater), Bayesian Essay Test Scoring sYstem (BETSY), Intelligent Essay Marking System (IEMS), Schema Extract Analyse and Report (SEAR), and The Essay Scoring Tool (TEST) [1]. The working techniques of few AEGs are discussed here.

PEG is one of the initial implementations of automated essay grading. PEG is a statistical approach based on the assumption that the quality of essays is reflected by the measurable proxes [2]. It uses factors such as “*proxes”* i.e. computer approximations or measures of *trins* which includes length of essay in terms of words to represent the trin of fluency; counts of prepositions, relative pronouns and other parts of speech. It also act as an indicator for complexity of sentence structure and variation in word length to indicate diction using previously manually marked essays as a training sets in order to calculate the regression coefficients. The other factor is in*trins*ic variables to simulate human rater grading. Natural Language Processing (NLP) technique and lexical content are not considered in PEG at all.

IEA is a domain-independent tool based on the Latent Semantic Analysis (LSA) technique that was originally designed for indexing documents and text retrieval [3]. LSA represents documents and their word content in a large two-dimensional matrix semantic space. Using a matrix algebra technique known as Singular Value Decomposition (SVD), new relationships between words and documents are uncovered, and existing relationship are modified to more accurately represent their true significance [4][5]. IEA includes relatively low unit cost, quick customized feedback, and plagiarism detection as its key features. The system is very well suited to analyze and score expository essays on science, social studies, history, medicine or business topics and automatically assesses and criticizes electronically submitted text essay [1].

E-Rater is a statistical and corpus based approach uses Microsoft Natural Language Processing tool for parsing the essay and to extract linguistic features from the essays and are finally evaluated against a benchmark set of human graded essays [6][7]. E-Rater includes domain based analysis of the discourse structure, of the syntactic structure and of the vocabulary usage. It is composed by five main independent modules. Three of these modules identify features for scoring guide criteria for the syntactic variety, the organization of ideas and the vocabulary usage of an essay. The rest modules are used to select and weigh predictive features for essay scoring and to compute the final score. A feedback component provide additional feedback about qualities of writing related to topic and fluency only

IEMS can be used both as an assessment tools and for diagnostic and tutoring purposes in many content-based subjects [8]. It is based on Pattern Indexing Neural Network (the Indextron). Indextron is defined as a specific clusterisation algorithm and can be implemented as a neural network embedded with an intelligent tutoring system for fast grading which provide feedback to students.

TEST is a domain based first AES tool for Hindi need prior knowledge before checking an essay. It uses quality of content, local coherence, factual accuracy, and global coherence as scoring parameters [9]. Each sentence in an essay is connected to previous sentences. The degree of this connection measures the coherence of the sentence pairs. Local coherence measures the inter sentence similarity whereas global coherence classify the structure of essays as good, average or bad. It takes human graded essays as training sets and rates them as good essays and bad essays. The fact evaluation module contain topic specific keywords, list of essays, correct facts list, and incorrect facts list and produce individual essay reports & scores with N X 1 Score Matrix for Internal use by TEST. It does not include grammatical checking and spell-check.

3. HINDI DEPENDENCY TREEBANK

Hindi Dependency Treebank (henceforth HDT) uses karaka - a syntactico-semantic relation as an intermediary step to express the semantic relations through vibhaktis [10]. Each karaka has a default vibhakti. In linguistics, grammatical relations (also called grammatical functions or grammatical roles, or syntactic functions) refer to functional relationships between constituents in a clause [11]. The role of grammatical relations in theories of grammar is greatest in dependency grammars, which tend to posit dozens of distinct grammatical relations. Every head-dependent dependency bears a grammatical function. Semantic analysis can be done using HDT as it includes Part-of-speech, Chunk Information, and Dependency Information. For each sentence, the output of HDT has four columns which are mentioned below,

- *1st Column* represents Token or chunk id such as 1, 1.1, 2, 2.2 etc.
- *2nd Column* indicates the actual word or word groups in the sentence having the attribute 'name' for naming.
- *3rd Column* specifies part of speech
- *4th Column* represents feature structure which holds morphological information, grammatical roles, semantic information etc..

4. LEXICAL ANALYSIS OF HINDI SENTENCES

A precise research in this decade has helped us to understand the AI & Machine Learning techniques based existing AEG systems, going through some limitations to propose a methodology which could work under Indian context. The methodology is based on the series of semantic evaluations.

For checking grammatical or semantic error, HDT of each sentence is captured. In this methodology, the features obtained from treebank are used to develop machine learning techniques to identify the errors. The machine learning procedure analyzes each noun, pronoun and verb and postposition associated with it. It also analyzes the number and gender agreement between noun/pronoun and verb.

Hindi is a free-word-order language but its default word order of sentences is Subject-Object-Verb (SOV). The object may be direct or indirect or both. In Hindi, postposition or vibhaktis/case marker is used instead of preposition in English language and is combined with noun or pronoun or more generally a noun phrase. Vibhaktis like ने,को ,से,का के,को,में, etc. are attached as suffix with noun or pronoun. Sentences in Hindi may follow default word order conventions for coding the information of grammatical relations. Hindi language has rich morphological case in which the subject and object and other verb arguments are identified in terms of the case markers that they bear (e.g. nominative, accusative, dative, genitive, ergative, etc.). The subject in a sentence must agree with the finite verb in person, number, and gender to be grammatical correct. A sentence is considered to be ungrammatical if it contains syntactic error. Let us consider the following sentences,

Eg1. राम ने रावण मारा.

Eg2. लड़की स्कूल में जाती हैं./ लड़की स्कूल को जाती हैं.

Eg3. गोपाल अपने भाई से लंबी हैं ।

Although Eg1. is ungrammatically as it is missing “को” after रावण in the sentence, HDT considers it to be grammatically correct as shown by the dependency structures of the sentences “राम ने रावण मारा”, “राम ने रावण को मारा” and “राम रावण को मारा” in the figure 1, 2 and 3 respectively where k1 indicates Karta karaka /Nominative Case (having ‘**ने**’ case marker) and k2 represent Karma karaka /Accusative case (having ‘**को**’ case marker).

This indicate that absence of case marker is not treated as grammatical error by HDT. Sentences in eg2. are grammatical but it could be more proper by eliminating the “में” and “को” as “लड़की स्कूल जाती हैं”. In eg3: गोपाल अपने भाई से लंबी हैं, the verb “लंबी” does not agree with the subject “गोपाल” as it possesses masculine gender whereas the verb here has feminine gender. Now consider the sentence “चला जाऐंगा अपने आप सुनील” which is ungrammatical too. Hence, the gender and number agreement helps in lexical analysis.

5. SEMANTIC ANALYSIS

Semantic knowledge provided information such as animacy, named entity categories and verb selectional restrictions. Named entity tag information is used to match the category of pronoun and their referent. The semantic class information (noun category) is used for the finding facts and fact evaluation in essays. The pairs which do not have semantic feature match are filtered out. Using the semantic knowledge for each word, semantic analysis is performed. "Semantic Analysis" refers to a formal analysis of meaning, and "computational" refer to approaches that in principle support effective implementation [12]. Semantic analysis involves the identification of the intended meaning at the word level i.e. word-sense disambiguation, as word has multiple meanings in different contexts. Semantic analysis also helps to understand that how different sentence and textual elements fit together. The analysis began with the identification of word senses computationally, exploring the interrelationships between the elements of a sentence, and relations between sentences (e.g., coreference), and examine the semantic relations and sentiment analysis. The dependency structures shown in figure 1,2, and 3 indicates that HDT shows the meaning of these sentences to be correct although they are grammatically incorrect. The dependency structure shows the relation of noun phrases and verb phrases which are semantically interrelated. Semantic knowledge analyzes multiple words and identifies their relations between as hypernymy & hyponymy and meronymy & holonymy too. Hindi WordNet is a system for bringing together different lexical and semantic relations between the Hindi words [13]. For each word (lexical item) there is a synonym set, or synset, in the Hindi WordNet, representing one lexical concept. Further, each synset is mapped to a concept ontology which defines the semantic properties of lexical items of a given synset.

Example

Word: फल

Possible Senses

Sense 1: Result

Related cluster snapshot: सफलता [success],द्धीप [island],फल [result], परिणाम [result],असफलता [failure],प्रततफल [failure]

Sense 2: Fruit

Related cluster snapshot: आम [mango],फल [fruit],भारत [India], खेल [game], मोटर [automobile]

Hence, the verbs and adverbs can be matched against the attributes related to various senses and shall manage the correlation between the segments of the sentences or clauses.

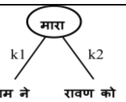

Figure 1.
Dependency Structure of
राम ने रावण को मारा

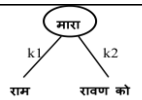

Figure 2.
Dependency Structure of
राम रावण को मारा

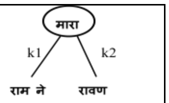

Figure 3.
Dependency Structure of
राम ने रावण मारा

6. CONCLUSION

The proposed methodology improves automated assessment by incorporating vast semantic attributes and grammar checking to overcome the issues related to automated essay evaluation systems. The system has to be evaluated on the basis of dependency and the supporting information from WordNet about sense and correctness of the sentences. In future, the size and variety of the corpus has to be increased. The factors of grammar checking other than number and gender agreements are considered as future research directions.

7. ACKNOWLEDGMENTS

This work was supported by Research and Development Laboratory, Department of Computer Science and Engineering at Bhilai Institute of Technology, Durg, Chhattisgarh, India, awaiting sponsorship from suitable funding agencies.

8. REFERENCES

[1] Salvatore Valenti, Francesca Neri and Alessandro Cucchiarelli. 2003. *An Overview of Current Research on Automated Essay Grading*. DIIGA - Universita' Politecnica delle Marche, Ancona, Italy Journal of Information Technology Education Volume 2

[2] Hearst, M. 2000. *The Debate On Automated Essay Grading*. IEEE Intelligent Systems, 15(5), 22-37

[3] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman R. A. 1990. *Indexing By Latent Semantic Analysis*. Journal of the American Society for Information Science, 41(6), 391-407.

[4] Whittington, D. & Hunt, H. 1999. *Approaches To The Computerized Assessment Of Free Text Responses*. In M. Danson (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough University, UK.

[5] Williams, R. 2001. *Automated Essay Grading: An Evaluation Of Four Conceptual Models*. In A. Hermann & M.M. Kulski (eds). Expanding Horizons in Teaching and Learning. Proceedings of the 10th Annual Teaching and Learning Forum, Perth: Curtin University of Technology.

[6] Burstein, J., Kukich, K., Wolff, S., Chi, L., & Chodorow M. 1998. *Enriching Automated Essay Scoring Using Discourse Marking*. Proceedings of the Workshop on Discourse Relations and Discourse

Marking, Annual Meeting of the Associationof Computational Linguistics, Montreal, Canada.

[7] Burstein, J., Leacock, C., & Swartz, R. 2001. *Automated Evaluation Of Essay And Short Answers*. In M. Danson (Ed.), Proceedingsof the Sixth International Computer Assisted Assessment Conference, Loughborough University, Loughborough, UK.

[8] Ming, P.Y., Mikhailov, A.A., & Kuan, T.L. 2000. *Intelligent Essay Marking System*. In C. Cheers (Ed.), Learners Together,Feb. 2000, NgeeANN Polytechnic, Singapore.

http://www.slideshare.net/singhg77/the-essay-scoring-tool-test-for-hindi

[9] Bharati, A., Sangal, R., Sharma, D.M., and Bai, L. 2006. *Anncorra: Annotating Corpora Guidelines For Pos And Chunk Annotation For Indian Languages*. In Technical Report (TRLTRC-31), LTRC, IIIT-Hyderabad.
https://en.wikipedia.org/wiki/Grammatical relation

[10] Blackburn, P., and Bos, J. 2005. *Representation and Inference For Natural Language: A First Course In Computational Semantics*, CSLI Publications. ISBN 1-57586-496-7.
http://www.cfilt.iitb.ac.in/wordnet/webhwn/